Model-Agnostic Covariate-Assisted Inference on Partially Identified Causal Effects

Wenlong Ji * Lihua Lei † Asher Spector *

September 12, 2024

Abstract

Many causal estimands are only partially identifiable since they depend on the unobservable joint distribution between potential outcomes. Stratification on pretreatment covariates can yield sharper partial identification bounds; however, unless the covariates are discrete with relatively small support, this approach typically requires consistent estimation of the conditional distributions of the potential outcomes given the covariates. Thus, existing approaches may fail under model misspecification or if consistency assumptions are violated. In this paper, we propose a unified and model-agnostic inferential approach for a wide class of partially identified estimands, based on duality theory for optimal transport problems. In randomized experiments, our approach can wrap around any estimates of the conditional distributions and provide uniformly valid inference, even if the initial estimates are arbitrarily inaccurate. Also, our approach is doubly robust in observational studies. Notably, this property allows analysts to use the multiplier bootstrap to select covariates and models without sacrificing validity even if the true model is not included. Furthermore, if the conditional distributions are estimated at semiparametric rates, our approach matches the performance of an oracle with perfect knowledge of the outcome model. Finally, we propose an efficient computational framework, enabling implementation on many practical problems in causal inference.

1 Introduction

1.1 Motivation and problem statement

Many parameters of interest in econometrics and causal inference are only *partially identifiable* (Manski, 2003; Tamer, 2010; Molinari, 2020a). Even in randomized experiments, we cannot observe the joint law of the potential outcomes $(Y_i(1), Y_i(0))$ since we observe at most one outcome per subject; thus, the law of the individual treatment effect $Y_i(1) - Y_i(0)$ is unidentifiable. However, most parameters of interest can be *bounded* using the marginal laws of $Y_i(1)$ and $Y_i(0)$. Furthermore, by incorporating information from covariates $X_i \in \mathbb{R}^p$, one can substantially reduce the width of the partially identified set.

However, partial identification bounds involving covariates can depend delicately on the relationship between the outcome and the covariates, making inference challenging. For illustration, we now give three motivating examples, although we will state a general problem formulation in Section 2. As notation, assume that we observe n i.i.d. observations $\{(X_i, W_i, Y_i)\}_{i=1}^n$ for covariates $X_i \in \mathcal{X}$, a binary treatment $W_i \in \{0, 1\}$ and an outcome $Y_i \in \mathcal{Y}$ with potential outcomes $Y_i(1), Y_i(0)$. This paper focuses on randomized experiments where the marginal laws of $(Y_i(1), X_i)$ and $(Y_i(0), X_i)$ are identified. Thus, we say that a parameter is *identified* if it depends only on these marginal laws.

Example 1 (Fréchet-Hoeffding bounds)**.** For fixed $y_1, y_0 \in \mathbb{R}$, let $\theta = \mathbb{P}(Y_i(1) \le y_1, Y_i(0) \le y_0)$ denote the joint CDF of the potential outcomes. θ is not identified but can be bounded. Indeed, without covariates,

*Department of Statistics, Stanford University

†Graduate School of Business, Stanford University

Authors are ordered alphabetically. We would like to thank P. M. Aronow, Bulat Garafov, Isaac Gibbs, Kevin Guo, Guido Imbens, Samir Khan, Yuichi Kitamura, Sokbae Lee, Dominik Rothenhäusler, Fredrik Sävje, Vira Semenova, and Jann Spiess for helpful discussion.

Hoeffding (1940); Fréchet (1951) showed that the sharp lower bound on θ is

$$\theta \geq \theta_L \coloneqq \max(0, \mathbb{P}(Y_i(1) \leq y_1) + \mathbb{P}(Y_i(0) \leq y_0) - 1). \tag{1}$$

With covariates, applying Eq. (1) conditional on X_i and integrating yields the sharp lower bound:

$$\theta \geq \theta_L \coloneqq \mathbb{E}\left[\max(0, \mathbb{P}(Y_i(1) \leq y_1 \mid X_i) + \mathbb{P}(Y_i(0) \leq y_0 \mid X_i) - 1)\right]. \tag{2}$$

Example 2 (Variance of the Individual Treatment Effect)**.** A natural measure of treatment effect heterogeneity is the variance of the individual treatment effect $\theta = \mathrm{Var}(Y_i(1) - Y_i(0))$. If θ is large relative to the average treatment effect, the treatment may be quite harmful for many individuals, and it is not clear if it should be given to the general population. The sharp lower bound on θ can be written as

$$\theta \geq \theta_L \coloneqq \mathrm{Var}\left(\mathbb{E}[Y_i(1) - Y_i(0) \mid X_i]\right) + \mathbb{E}[\mathrm{Var}_{U \sim \mathrm{Unif}(0,1)}(P^{\star\,-1}_{Y(1)|X}(U \mid X_i) - P^{\star\,-1}_{Y(0)|X}(U \mid X_i))] \tag{3}$$

where $P^{\star}_{Y(k)|X}$ denotes the true conditional CDF of $Y_i(k) \mid X_i$ for $k \in \{0, 1\}$.

Example 3 (Average treatment effect with selection bias)**.** Suppose in a randomized experiment, we only observe outcomes for a set of "selected" individuals, where selection may depend on treatment status. E.g., a canonical example is that we only observe wages for individuals who are employed (Lee, 2009), but treatment may affect employment. Formally, let $S_i \in \{0, 1\}$ be the indicator for the selection event, with $S_i(1), S_i(0)$ its potential outcomes. A natural estimand in this context is the ATE for the individuals who would be selected with or without the treatment:

$$\theta \coloneqq \mathbb{E}[Y_i(1) - Y_i(0) \mid S_i(1) = S_i(0) = 1]. \tag{4}$$

θ is only partially identifiable, but as in Example 2, if we can learn the relationship between Y_i, S_i and X_i, then we give sharp bounds on θ. In particular, Semenova (2021) showed that if one assumes that selection is "monotone" in the treatment, meaning $S_i(1) \geq S_i(0)$ a.s., then the sharp lower bound is

$$\theta \geq \theta_L \coloneqq \mathbb{E}_X[\mathbb{E}[Y_i(1) | S_i(1) = 1, X_i, Y_i(1) \leq Q_{\eta(X_i)}(X_i)]] - \mathbb{E}[Y_i(0) | S_i(0) = 1] \tag{5}$$

where above, $\eta(X_i) \coloneqq \frac{\mathbb{P}(S_i(0)=1|X_i)}{\mathbb{P}(S_i(1)=1|X_i)}$ and $Q_\alpha(X_i)$ denotes the α conditional quantile of $Y_i(1) \mid X_i$. These bounds are colloquially known as "Lee bounds."[1] More general problems of this sort are studied in the literature on principal stratification (e.g. Frangakis and Rubin, 2002; Page et al., 2015; Ding and Lu, 2017).

Given a partially identified parameter θ, this paper aims to estimate sharp bounds $[\theta_L, \theta_U]$ which incorporate information from covariates. This problem is challenging because the bounds typically depend delicately on the conditional law of $Y \mid X$, as exemplified by Equations (3)-(5). Thus, most existing approaches to estimate θ_L, θ_U assume that one can uniformly consistently estimate such nuisance parameters (see Section 1.3 for a review). This assumption is often implausible when X_i is continuous or categorical with many values or high-dimensional, unless the researcher is willing to impose further assumptions on the conditional distributions (e.g., a parametric model, smoothness, sparsity), which may not hold in applications.

Thus, in this work, we ask the question: can we obtain valid conservative inference on the sharp identified set $[\theta_L, \theta_U]$ without making *any* assumptions about the conditional distributions? Furthermore, can we do this in a way that is asymptotically efficient given consistent estimates of the conditional distributions?

We end this section by noting that this question is motivated by the core philosophy of partial identification. Indeed, why not simply make enough assumptions so that the parameter θ is identified? In his seminal book, Manski (2003) answers this question by formulating the law of decreasing credibility:

> The credibility of inference decreases with the strength of the assumptions maintained.

Our objective is to enhance *credibility* by removing any assumptions about the true outcome model or the accuracy of the researcher's working model, without sacrificing *efficiency* when the researcher's working model matches the ground truth.

[1] The choice of terminology pays tribute to Lee (2009), although the initial derivation of the sharp bounds without covariates can be attributed to Zhang and Rubin (2003).

1.2 Contribution and overview of results

Our work introduces a framework for inference on sharp, covariate-assisted partial identification bounds on causal parameters. In particular, if $P^\star$ denotes the true joint law of $(Y_i(1), Y_i(0), X_i) \overset{\text{i.i.d.}}{\sim} P^\star$, we consider estimands of the form

$$\theta(P^\star) \coloneqq \mathbb{E}_{P^\star}\left[f(Y(0), Y(1), X)\right] \tag{6}$$

for some known function $f : \mathcal{Y}^2 \times \mathcal{X} \to \mathbb{R}$. Many estimands of interest can be reduced to this case, including Examples 1-3, certain conditional expectations, quantiles of treatment effects, and more (see Section 2.5). We let $\theta_L \leq \theta(P^\star) \leq \theta_U$ denote the sharp (population) lower and upper partial identification bounds on $\theta(P^\star)$; these quantities are defined formally in Section 2.1.

Our method outputs estimates $\hat{\theta}_L, \hat{\theta}_U$ of the sharp bounds θ_L, θ_U as well as lower and upper confidence bounds $\hat{\theta}_{\text{LCB}}, \hat{\theta}_{\text{UCB}}$. The main idea is to leverage duality theory for optimal transport problems (reviewed in Section 2.1) to convert any estimate $\hat{P}_{Y|X,W}$ of the conditional law of the outcome into robust partial identification bounds $\hat{\theta}_L, \hat{\theta}_U$. We emphasize that this method works automatically for any estimand defined above—in our software, the analyst can specify any function f and does not need to do any additional calculations to obtain the results. This eliminates the need to obtain a closed-form representation for the sharp bounds θ_L, θ_U.

These "dual bounds" have a few appealing properties, listed below.

1. Uniform validity. Our method allows analysts to estimate the law of $Y \mid X, W$ using any machine learning technique, e.g., quantile regression, boosting, neural networks, etc. However, in randomized experiments with known propensity scores, the resulting confidence bounds are valid *even if* the machine-learning estimate $\hat{P}_{Y|X,W}$ is arbitrarily inaccurate relative to the ground truth $P^\star_{Y|X,W}$. In this sense, our method is "model-agnostic": it can leverage models for power, but does not rely on them for validity.

Formally, $\hat{\theta}_L$ and $\hat{\theta}_U$ are always *conservatively* biased in the sense that $\mathbb{E}[\hat{\theta}_L] \leq \theta_L$ and $\mathbb{E}[\hat{\theta}_U] \geq \theta_U$. Furthermore, the confidence bounds have asymptotic coverage, e.g.,

$$\liminf_{n\to\infty} \mathbb{P}(\hat{\theta}_{\text{LCB}} \leq \theta_L) \geq 1 - \alpha, \tag{7}$$

without any assumptions on the accuracy of $\hat{P}_{Y|X,W}$. This result can easily be extended to a uniform validity result (see Theorem 3.1). Finally our method is also doubly robust in observational studies where the propensity scores are not known (see Theorem 3.5).

2. Tightness. If one can estimate the relevant features of the true conditional law $P^\star_{Y|X,W}$ at $o(n^{-1/4})$ rates, our estimators $\hat{\theta}_L, \hat{\theta}_U$ are asymptotically indistinguishable from oracle unbiased estimators which have full knowledge of all unknown nuisance parameters. This implies that the estimates $\hat{\theta}_L, \hat{\theta}_U$ are asymptotically unbiased and $\sqrt{n}$-consistent for the sharp bounds θ_L, θ_U.

3. Easy model selection. A major question in empirical applications is (i) how to select the subset of the covariates used in the analysis and (ii) how to estimate the outcome model $Y_i \mid X_i, W_i$. Our method permits the analyst to use either nested cross-validation and or the multiplier bootstrap (Chernozhukov et al., 2013a) to select the tightest bound based on different models or subsets of the covariates.

4. Computational efficiency. To compute our bounds, we propose an algorithm which is computationally efficient even when X_i is high-dimensional and does not require Y_i to be discrete or bounded. The python package `dualbounds` implements this algorithm: `https://dualbounds.readthedocs.io/en/latest/`.

Figure 1 illustrates our contributions in a simple numerical experiment where we estimate lower Lee bounds as in Example 3. We fit an outcome model estimate $\hat{P}_{Y|X,W}$ assuming $Y_i(k) \mid X_i$ follows a homoskedastic Gaussian linear model for $k \in \{0, 1\}$. A naive estimator of θ_L, which simply plugs in the estimated outcome model to Equation (5), performs very well when the model is well-specified. However, in contrast, if the errors are made heteroskedastic, this naive "plug-in" estimator can become conservatively or anticonservatively biased (depending on the form of heteroskedasticity). In contrast, our dual bounds wrap around exactly the same estimator of the outcome model and provide provable validity under arbitrary misspecification. Indeed, Figure 1 shows that dual bounds perform nearly as well as a "dual oracle" method that has perfect knowledge of the true heterskedasticity pattern. See Section 5.4 for precise simulation details and an analogous plot showing coverage.

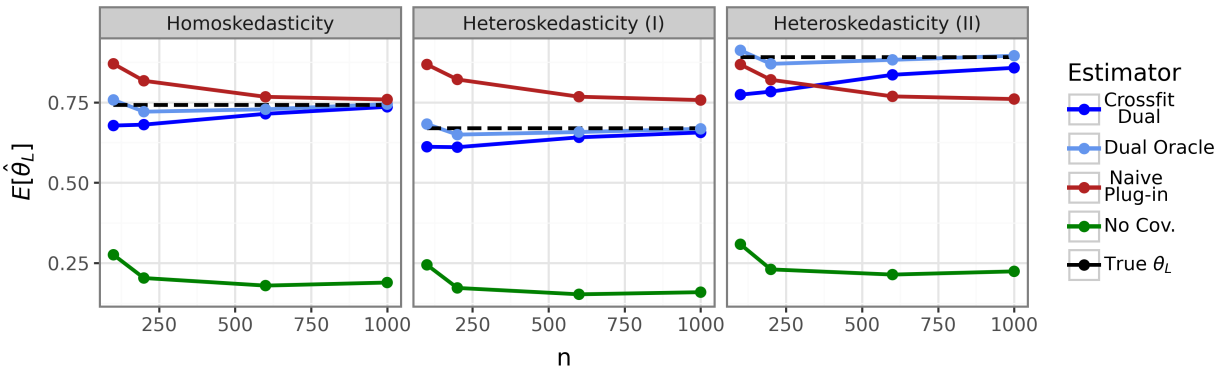

Figure 1: This figure illustrates our core contribution in a simple setting where we aim to estimate lower Lee bounds (Example 3). For each method, it shows the average value of the estimate of θ_L, and the dotted black line shows the true value of the sharp lower bound θ_L. The green covariate-free approach is highly conservative. However, under misspecification (in this case, two forms of heteroskedasticity), a naive covariate-assisted plug-in estimator can become conservative or anti-conservative. In contrast, our dual bound methodology is (i) at worst conservative under misspecification and (ii) competitive with the provably unbiased "dual oracle" estimator. See Section 5.4 for precise simulation details.

1.3 Related literature

Partial identification has a long history in econometrics and causal inference, and a great deal of work has been done to characterize and estimate sharp bounds in various settings (e.g. Manski, 1990, 1997; Balke and Pearl, 1997; Heckman et al., 1997; Manski and Tamer, 2002; Imbens and Manski, 2004; Firpo and Ridder, 2008; Molinari, 2008; Lee, 2009; Stoye, 2009; Fan and Park, 2010; Chiburis, 2010; Fan and Park, 2012; Tetenov, 2012; Andrews and Shi, 2013; Aronow et al., 2014; Fan et al., 2017; Firpo and Ridder, 2019; Russell, 2021; Jun and Lee, 2023; Ober-Reynolds, 2023); see Manski (2003); Tamer (2010); Molinari (2020b) for a review. When covariates are available, the bounds can be improved by conditioning on the covariates and aggregating covariate-specific sharp bounds (Chernozhukov et al., 2007; Chandrasekhar et al., 2012; Chernozhukov et al., 2013c; Semenova and Chernozhukov, 2020; Semenova, 2021, 2023; Lee, 2023; Levis et al., 2023). However, unless the covariates are discrete with a few values, these methods generally either (a) make assumptions that allow the conditional distributions of the potential outcomes to be consistently estimated at semiparametric rates or (b) have to discretize covariates in a non-disciplined way at the cost of efficiency loss. In contrast, our method can handle any type of covariates without making any assumptions that enable consistent estimates of the conditional distributions.

Our key technical tool is the theory of duality in optimization. This tool is of course not new, although we use it in a novel way. In particular, many existing works use duality theory as part of an inference strategy, for example in analysis of certain linear programming problems (e.g., Hsieh et al. (2022); Andrews et al. (2023); Fang et al. (2023)) and in sensitivity analysis (Dorn and Guo, 2022; Dorn et al., 2022). Most recently, Semenova (2023) independently developed a dual-based estimator for a class of intersection bounds (Chernozhukov et al., 2013c). However, it is unclear whether it can be applied to causal estimands considered in this paper while achieving the same level of robustness. In fact, they require consistent estimates of the conditional distributions at semiparametric rates uniformly over the covariate space. Moreover, their method needs to calculate all vertices of the dual constraint polytope, which is only possible for discrete potential outcomes and in general requires exponential time in the support size of the potential outcomes.

2 Core Methodology

Note that to aid comprehension, we will mostly defer measure-theoretic details to Appendix C. We defer all computational details to Section 4.

2.1 Notation and background on Kantorovich duality

We begin by slightly generalizing the setting introduced in Section 1. In particular, we allow the analyst to specify a *model class* $\mathcal{P}$ of possible probability distributions over $\mathcal{Y}^2 \times \mathcal{X}$. By default, we recommend taking $\mathcal{P}$ to be the unrestricted class of all probability distributions over $\mathcal{Y}^2 \times \mathcal{X}$; however, in some settings, analysts may prefer to incorporate a-priori additional assumptions about the law of $(Y_i(0), Y_i(1), X_i)$. We allow analysts to do this using a collection of conditional moment inequalities, as defined below:

Assumption 2.1. *For each $x \in \mathcal{X}$, let $\mathcal{W}_x = \{w_{x,1}, \ldots, w_{x,L}\}$ denote a finite collection of functions mapping $\mathcal{Y}^2 \to \mathbb{R}$ for $L \in \mathbb{N}$.*[2] *Suppose $\mathcal{P}$ takes the form*

$$\mathcal{P} = \Big\{ \textit{joint distributions } P \textit{ over } \mathcal{Y}^2 \times \mathcal{X} \text{ s.t. } \mathbb{E}_P[w(Y(0), Y(1)) \mid X = x] \leq 0 \ \ \forall w \in \mathcal{W}_x, x \in \mathcal{X} \Big\}.$$

For example, $\mathcal{P}$ is simply the set of all joint distributions over $\mathcal{Y}^2 \times \mathcal{X}$ if $\mathcal{W}_x$ is empty for each $x \in \mathcal{X}$. On the other hand, in the setting of Lee bounds (Example 3) with compound potential outcomes $(Y_i(0), S_i(0)), (Y_i(1), S_i(1))$, the monotonicity assumption in Lee (2009) can be enforced by setting $\mathcal{W}_x$ to contain the single function $w(y_0, s_0, y_1, s_0) = \mathbb{I}(s_0 > s_1)$, which ensures $S_i(0) \leq S_i(1)$ a.s. The conditional monotonicity assumption of Semenova (2021) also satisfies Assumption 2.1.

Given $\mathcal{P}$, the lower bound θ_L is the minimum value of $\theta(P)$ for all $P \in \mathcal{P}$ which are consistent with the true marginal distributions $P^\star_{Y(1),X}$ and $P^\star_{Y(0),X}$:

$$\theta_L = \inf_{P \in \mathcal{P}} \mathbb{E}_P[f(Y(0), Y(1), X)] \text{ s.t. } P_{Y(1),X} = P^\star_{Y(1),X} \text{ and } P_{Y(0),X} = P^\star_{Y(0),X}.$$

We assume throughout that the true law $P^\star$ is an element of the model class $\mathcal{P}$. This assumption is automatically satisfied if one takes $\mathcal{P}$ to be the unrestricted model class (which we recommend).

Now, we introduce the dual to the optimization problem defining θ_L. Namely, for *dual variables* $\nu_0, \nu_1 : \mathcal{Y} \times \mathcal{X} \to \mathbb{R}$, the Kantorovich dual function is

$$g(\nu_0, \nu_1) \coloneqq \mathbb{E}_{P^\star}[\nu_0(Y(0), X) + \nu_1(Y(1), X)]. \tag{8}$$

We aim to use $g(\nu_0, \nu_1)$ as a lower bound on θ_L. To ensure $g(\nu_0, \nu_1) \leq \theta_L$ holds, we will enforce a collection of known (if complex) constraints on ν_0, ν_1. For example, in the simplest case where $\mathcal{P}$ is unrestricted, we can simply require that $\nu_0(y_0, x) + \nu_1(y_1, x) \leq f(y_0, y_1, x)$ for all $y_1, y_0, x \in \mathcal{Y}^2 \times \mathcal{X}$ as discussed in Section 1. In the general case where $\mathcal{W}_x = \{w_{x,1}, \ldots, w_{x,L}\}$ is nonempty, we can slightly loosen these constraints to take advantage of additional assumptions on $\mathcal{P}$.

In particular, for any $\nu_k : \mathcal{Y} \times \mathcal{X} \to \mathbb{R}$, let $\nu_{k,x} : \mathcal{Y} \to \mathbb{R}$ denote the function $\nu_{k,x}(y) \coloneqq \nu_k(y, x)$ which holds the second coordinate of ν_k constant. For $x \in \mathcal{X}$, we say that $\nu_{0,x}, \nu_{1,x}$ are *conditionally valid* at x if there are a collection of nonnegative constants $\{\lambda_{x,\ell}\}_{\ell=1}^L$ such that the following holds:

$$\nu_{0,x}(y_0) + \nu_{1,x}(y_1) \leq f(y_0, y_1, x) + \sum_{\ell=1}^{L} \lambda_{x,\ell} \cdot w_{x,\ell}(y_0, y_1) \qquad \text{for all } y_0, y_1 \in \mathcal{Y} \tag{9}$$

and we let $\mathcal{V}_x \subset \{\mathcal{Y} \to \mathbb{R}^2\}$ denote the set of all pairs of functions satisfying this condition. Finally, we say that ν_0, ν_1 are fully valid or "dual-feasible" if $\nu_{0,x}, \nu_{1,x} \in \mathcal{V}_x$ are conditionally valid for every $x \in \mathcal{X}$, and we let $\mathcal{V} \subset \{\mathcal{Y} \times \mathcal{X} \to \mathbb{R}^2\}$ denote the set of all valid dual variables.

Computational issues aside (see Section 4), we emphasize that $\mathcal{V}$ is a *known* set which does not depend on $P^\star$. Crucially, satisfying this known constraint ensures that weak duality holds, e.g., $g(\nu_0, \nu_1) \leq \theta_L$. The theorem below states this formally; it also states that strong duality holds and gives a useful characterization of the optimal dual variables $\nu^\star \in \arg\max_{\nu \in \mathcal{V}} g(\nu)$ which we will use throughout the paper. To ease readability, we defer technical regularity conditions regarding measurability to Appendix C.

Theorem 2.1 (Kantorovich duality)**.** *Under Assumption 2.1, the following holds:*

1. Weak duality*: For any $(\nu_0, \nu_1) \in \mathcal{V}$, $g(\nu_0, \nu_1) \leq \theta_L$.*
2. Strong duality*: Under mild measurability conditions and regularity conditions on f and $\mathcal{W}_x$ stated in Appendix C, there exist $(\nu_0^\star, \nu_1^\star) \in \mathcal{V}$ such that $g(\nu_0^\star, \nu_1^\star) = \theta_L$. Furthermore, for each $x \in \mathcal{X}$,*

$$\nu^\star_{0,x}, \nu^\star_{1,x} \in \underset{\nu_{0,x}, \nu_{1,x} \in \mathcal{V}_x}{\arg\max} \ \mathbb{E}_{P^\star_{Y(0)|X=x}}[\nu_{0,x}(Y(0))] + \mathbb{E}_{P^\star_{Y(1)|X=x}}[\nu_{1,x}(Y(1))]. \tag{10}$$

[2] Our theory allows $|\mathcal{W}_x| = L$ to vary with x but for simplicity our notation suppresses this dependence.

Note that a proof of Theorem 2.1 can be found in Appendix C, and a discussion of the regularity conditions on our examples is presented in Remark 15; our proof uses techniques from Zaev (2015).

This theorem has two statistical implications. First, to estimate θ_L, we need only estimate $\nu^\star$. Second, $\nu^\star$ is only a functional of $P^\star_{Y(0)|X}, P^\star_{Y(1)|X}$ and does not depend on $P^\star_X$. We now use these insights to devise a strategy to estimate θ_L.

2.2 Inference via dual bounds

To motivate our method, note the dual function $g(\nu)$ is easy to estimate for any fixed choice of $\nu \in \mathcal{V}$ using (e.g.) an IPW estimator. Thus, if only we knew the value of $\nu^\star$, we could easily estimate $\theta_L = g(\nu^\star)$. The main idea is to use the first split of the data to estimate $\hat{\nu} \approx \nu^\star$, and the second split of the data to estimate $g(\hat{\nu})$. Crucially, even if our first-stage estimate $\hat{\nu}$ is poor, our inference will be conservative but valid, since weak duality ensures $g(\hat{\nu}) \leq \theta_L$. And as we will see in Section 3, if $\hat{\nu}$ is close to $\nu^\star$, then our inference will be very tight.

Definition 1 (Dual lower bounds)**.** Given data $\{(Y_i, W_i, X_i)\}_{i=1}^n$, we first randomly split the data into two disjoint subsets $\mathcal{D}_1$ and $\mathcal{D}_2$ of equal size. Then we perform the following steps:

Step 1: On $\mathcal{D}_1$, compute any estimator $\hat{\nu} \in \mathcal{V}$ for $\nu^\star \coloneqq \arg\max_{\nu \in \mathcal{V}} g(\nu)$. There are many reasonable ways to do this, but we suggest the following method:

(a) Step 1a: Compute an estimate $\hat{P}_{Y(0)|X}, \hat{P}_{Y(1)|X}$ of the conditional laws $P^\star_{Y(0)|X}, P^\star_{Y(1)|X}$. To do this, one can use any machine-learning or regression algorithm, such as lasso-based techniques, regularized quantile regression, or distributional regression (Kneib et al., 2023)—see Section 2.4 for more details.

(b) Step 1b: Let $\hat{\nu}$ maximize the "empirical dual" $\hat{g}$ which plugs in $\hat{P}_{Y(0)|X}, \hat{P}_{Y(1)|X}$ for $P^\star$. In particular, to define $\hat{\nu}$, we use the characterization from Theorem 2.1. Precisely, for each $x \in \mathcal{X}$, define the functions $\hat{\nu}_{0,x}, \hat{\nu}_{1,x} : \mathcal{Y} \to \mathbb{R}$ as the solution to

$$\hat{\nu}_{0,x}, \hat{\nu}_{1,x} = \underset{\nu_{0,x}, \nu_{1,x} \in \mathcal{V}_x}{\arg\max} \mathbb{E}_{\hat{P}_{Y(0)|X=x}}[\nu_{0,x}(Y(0))] + \mathbb{E}_{\hat{P}_{Y(1)|X=x}}[\nu_{1,x}(Y(1))]. \tag{11}$$

Then, we define the full dual variables $\hat{\nu}_0, \hat{\nu}_1 : \mathcal{Y} \times \mathcal{X} \to \mathbb{R}$ such that

$$\hat{\nu}_0(y, x) \coloneqq \hat{\nu}_{0,x}(y) \text{ and } \hat{\nu}_1(y, x) \coloneqq \hat{\nu}_{1,x}(y) \text{ for all } y, x \in \mathcal{Y} \times \mathcal{X}. \tag{12}$$

We note that Eq. (11) may not have a unique solution, in which case we suggest taking the minimum norm solution to reduce the variance of the final estimator—see Appendix A.1 for details. Computing $\hat{\nu}$ may seem challenging, but we will discuss simple methods to do this in Section 4. For now, we merely note our the final estimator depends only on $\hat{\nu}_{0,x}, \hat{\nu}_{1,x}$ for $x \in \{X_i : i \in \mathcal{D}_2\}$ and thus we do not need to solve Eq. (11) for all $x \in \mathcal{X}$.

Step 2: Define $\tilde{\theta}_L \coloneqq g(\hat{\nu})$, and note by weak duality that $\tilde{\theta}_L \leq \theta_L$ holds deterministically. On $\mathcal{D}_2$, we will define a conservative estimator of θ_L by using an IPW estimator which is unbiased for $\tilde{\theta}_L$. In particular, define

$$\hat{\theta}_L \coloneqq \frac{1}{|\mathcal{D}_2|} \sum_{i \in \mathcal{D}_2} \frac{\hat{\nu}_1(Y_i, X_i) W_i}{\pi(X_i)} + \frac{\hat{\nu}_0(Y_i, X_i)(1 - W_i)}{1 - \pi(X_i)}. \tag{13}$$

Conditional on $\mathcal{D}_1$, $\hat{\theta}_L$ equals a sample mean of i.i.d. terms, and $\hat{\theta}_L$ is conservatively biased for θ_L. Thus, we can compute a lower confidence bound on θ_L via the univariate central limit theorem. In particular, let $\hat{\sigma}_S$ denote the sample standard deviation of the summands $\left\{\frac{\hat{\nu}_1(Y_i, X_i) W_i}{\pi(X_i)} + \frac{\hat{\nu}_0(Y_i, X_i)(1 - W_i)}{1 - \pi(X_i)}\right\}_{i \in \mathcal{D}_2}$. Then a $1 - \alpha$ lower confidence bound (LCB) for θ_L is

$$\hat{\theta}_{\text{LCB}} = \hat{\theta}_L - \Phi^{-1}(1 - \alpha) \frac{\hat{\sigma}_S}{\sqrt{|\mathcal{D}_2|}} \tag{14}$$

where Φ is the standard Gaussian CDF.

We will see in Section 3 that this procedure is always valid (in randomized experiments) and that it is "efficient" if we can estimate $\hat{P}_{Y(0)|X}, \hat{P}_{Y(1)|X}$ at semiparametric rates. The main drawbacks of this procedure are that it requires splitting the data and that Eq. (13) assumes the propensity scores are known. In Section 3, we partially overcome these drawbacks by employing cross-fitting and by plugging in estimates $\hat{\pi}$ of the propensity scores in observational data. In most (but not all) settings, these tactics will preserve validity and increase efficiency. Before presenting our theoretical results, however, we first give a few guidelines and examples of how to apply this procedure.

2.3 Model selection via the multiplier bootstrap

To compute dual bounds, one must estimate the dual variables $\nu^\star$—in practice, we recommend first estimating $\hat{P}_{Y(0)|X}, \hat{P}_{Y(1)|X}$ and then computing $\hat{\nu}$ as per Eq. (12). However, there are many ways to estimate $\hat{P}_{Y(0)|X}, \hat{P}_{Y(1)|X}$. E.g., analysts may prefer to use only a subset of the covariates to predict Y, but even after observing $\mathcal{D}_1$, it is not clear which subset of the covariates to choose. And even after making this decision, as discussed in Section 2.4, there are still countless existing methods to estimate $\hat{P}_{Y(0)|X}, \hat{P}_{Y(1)|X}$. This raises the question: in practice, how should analysts choose between K candidate estimates $\hat{\nu}^{(1)}, \ldots, \hat{\nu}^{(K)} \in \mathcal{V}$ of the dual variables $\nu^\star$? Or more colloquially, how should we perform model selection?

One solution is to perform cross-validation within the first fold ($\mathcal{D}_1$) and pick the best-performing model. This approach is clearly valid since the final estimated dual variables $\hat{\nu}$ still depend only on $\mathcal{D}_1$, satisfying Definition 1. In Section 3.4, we recommend this approach for observational studies, where the validity of the final bounds may depend on the accuracy of the outcome model. However, in randomized experiments, we can substantially improve upon this method.

In particular, let $\tilde{\theta}_L^{(k)} = g(\hat{\nu}^{(k)})$ denote the dual lower bound on θ_L implied by the estimate $\hat{\nu}^{(k)}$, for $k = 1, \ldots, K$. We will estimate $\max_{k\in[K]} \tilde{\theta}_L^{(k)}$, the tightest possible lower bound on θ_L based on $\{\hat{\nu}^{(k)}\}_{k=1}^K$, using the Gaussian multiplier bootstrap (Chernozhukov et al., 2013a), as defined below.

Definition 2 (Dual bounds with the multiplier-bootstrap)**.** Given dual variables $\hat{\nu}^{(1)}, \ldots, \hat{\nu}^{(K)} \in \mathcal{V}$, for $i \in \mathcal{D}_2$, define the IPW summands as:

$$S_i^{(k)} \coloneqq \frac{\hat{\nu}_1^{(k)}(Y_i, X_i)W_i}{\pi(X_i)} + \frac{\hat{\nu}_0^{(k)}(Y_i, X_i)(1 - W_i)}{1 - \pi(X_i)} \text{ for } k \in [K]. \tag{15}$$

Define $\hat{\theta}_L^{(k)} = \frac{1}{|\mathcal{D}_2|}\sum_{i\in\mathcal{D}_2} S_i^{(k)}$ and $\hat{\sigma}_k^2 = \frac{1}{|\mathcal{D}_2|}\sum_{i\in\mathcal{D}_2}(S_i^{(k)} - \hat{\theta}_L^{(k)})^2$ to be the dual estimators and associated sample variances for each $k \in [K]$. The main idea is to use $T \coloneqq \max_{k\in[K]} \frac{\sqrt{n}\hat{\theta}_L^{(k)}}{\hat{\sigma}_k}$ as a test statistic and compute its quantile using the Gaussian multiplier bootstrap. Precisely:

1. Sample $W_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$ for each $i \in \mathcal{D}_2$.
2. Let $T^{(b)} = \max_{k\in[K]} \hat{\sigma}_k^{-1}\left[\frac{1}{\sqrt{|\mathcal{D}_2|}}\sum_{i\in\mathcal{D}_2} W_i(S_i^{(k)} - \hat{\theta}_L^{(k)})\right]$ be the bootstrapped test statistic.
3. Let $\hat{q}_{1-\alpha} \coloneqq Q_{1-\alpha}(T^{(b)} \mid \mathcal{D})$ be the $1-\alpha$ quantile of $T^{(b)}$ conditional on the data. This can be computed by simulating many bootstrap samples.

Then, return the following multiplier bootstrap (MB) lower confidence bound:

$$\hat{\theta}_{\text{LCB}}^{\text{MB}} \coloneqq \max_{k\in[K]} \left\{\hat{\theta}_L^{(k)} - \hat{q}_{1-\alpha}\frac{\hat{\sigma}_k}{\sqrt{|\mathcal{D}_2|}}\right\}. \tag{16}$$

The multiplier bootstrap is well-suited to this problem for two reasons. First, our bounds are valid no matter which model we select, i.e., $\tilde{\theta}_L^{(k)} \le \theta_L$ always holds. This may not be true in other problems—for example, when estimating regression coefficients, selecting different subsets of covariates may lead to anticonservative bias, but in our setting, any bias from misspecification is conservative. Second, after estimating $\{\hat{\nu}^{(k)}\}_{k=1}^K$, the dual bounds $\{\tilde{\theta}_L^{(k)}\}_{k=1}^K$ can be expressed as marginal moments and estimating them does not require (e.g.) any complicated M-estimation. As a result, in Section 3.1, we conclude that the multiplier bootstrap quantile $\hat{q}_{1-\alpha}$ is consistent even if K grows exponentially with a power of n (Chernozhukov et al., 2018b).

Remark 1. If K is too large, we note that $\hat{\theta}_{\text{LCB}}^{\text{MB}}$ may be conservative since the quantile $\hat{q}_{1-\alpha}$ will grow with K. One solution is to use tools from selective inference (e.g. Andrews et al., 2019; Zrnic and Fithian, 2022). We leave this possibility to future work.

Remark 2. In some of our empirical applications (Section 5), the estimands can only be expressed as the *ratio* of two marginal moments. We can extend the multiplier bootstrap methodology to that setting under the restriction that K cannot grow with n. For brevity, we present this extension in Appendix A.2.

2.4 Guidelines on estimating the conditional distributions $\hat{P}_{Y(0)|X}, \hat{P}_{Y(1)|X}$

The first step in computing a dual bound $\hat{\theta}_{\text{LCB}}$ is to estimate $\hat{P}_{Y(0)|X}, \hat{P}_{Y(1)|X}$, or equivalently (in randomized experiments), to estimate the conditional law of $Y_i \mid X_i, W_i$. An immense literature exists on this modeling

problem (e.g. Koenker and Bassett Jr, 1978; Chernozhukov et al., 2010, 2013b; Friedman, 2020), and any choice of method will be valid regardless of misspecification. However, we make a few recommendations here.

To start, note that it is not usually sufficient to model the *conditional mean* $\mathbb{E}_{P^\star}[Y_i \mid X_i, W_i]$, since the sharp lower bound θ_L may depend on the whole conditional law (e.g. Examples 1- 3). The good news is that an entire field called *distributional regression* is devoted to the task of estimating the law $Y_i \mid X_i, W_i$ (see Kneib et al. (2023) for a review). One way to do this is to fit many quantile regressions. Another simple method is to assume a Gaussian linear model, i.e.,

$$Y_i = \phi(X_i, W_i)^T \beta + \epsilon_i \tag{17}$$

where $\phi(X_i, W_i) \in \mathbb{R}^d$ is some feature transformation of X_i, W_i and $\epsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$. To fit this model, one can (i) adaptively fit the feature representation ϕ using the first fold $\mathcal{D}_1$, (ii) fit a regularized estimate $\hat{\beta}$ of β using (e.g.) a cross-validated lasso on $\mathcal{D}_1$, and (iii) estimate σ^2 using the usual OLS estimator of the residual variance. Of course, the Gaussian assumption may not always be realistic. Instead, our default implementation in `dualbounds` fits the same coefficients $\hat{\beta}$ and uses the empirical residuals $\hat{\epsilon}_i := Y_i - \phi(X_i, W_i)^T \hat{\beta}$ to nonparametrically estimate the law of ϵ_i. Similarly, in the presence of heteroskedasticity, we can estimate $\text{Var}(Y_i \mid X_i, W_i)$ using a nonparametric estimator like a random forest; clearly, the possibilities are endless. The main point is that misspecification of these models will not affect the validity of $\hat{\theta}_{\text{LCB}}$, although better models will yield tighter confidence intervals.

2.5 Examples

In this section, we give a few examples of estimands that fit into the framework from Section 2.2. We start with a few easy examples before returning to Lee bounds (Example 3), which are slightly more complicated.

Example 1 (Fréchet-Hoeffding bounds)**.** The joint CDF of the potential outcomes evaluated at a fixed point $(y_1, y_0) \in \mathcal{Y}^2$ is clearly an expectation over P, i.e., $\theta(P) := \mathbb{E}_P[\mathbb{I}(Y_i(1) < y_1, Y_i(0) < y_0)]$.

Example 2 (Variance of the ITE)**.** If $\theta(P) = \text{Var}_P(Y_i(1) - Y_i(0))$, we can write

$$\theta(P) = \mathbb{E}_P[(Y_i(1) - Y_i(0))^2] - (\mathbb{E}_P[Y_i(1) - Y_i(0)])^2 .$$

Note that the left-hand term is an expectation over P, and the right-hand term is perfectly identifiable: it is just the square of the average treatment effect. Thus, we can apply the dual-bound methodology to the left-hand term, and we can estimate the right-hand term using another IPW estimator on $\mathcal{D}_2$. The only adjustment from Definition 1 is that we use the bivariate delta method to compute the appropriate standard errors for θ_L (see Appendix A.2 for a full derivation).

Example 4 (Makarov bounds)**.** Define $\theta(P) := \mathbb{E}_P[\mathbb{I}(Y_i(1) - Y_i(0) < t)]$ to be the CDF of the ITE at a fixed point $t \in \mathbb{R}$. Again, $\theta(P)$ is clearly an expectation over P.

We now return to the case of Lee bounds (Example 3) from Section 1.1.

Example 3 (Lee bounds)**.** Suppose $\theta(P) := \mathbb{E}[Y_i(1) - Y_i(0) \mid S_i(1) = S_i(0) = 1]$ is the ATE for the "always takers," i.e., the subset of individuals who would be selected under treatment or control. In this problem, we have bivariate potential outcomes of the form $(Y_i(0), S_i(0))$ and $(Y_i(1), S_i(1))$, which differs slightly from the notation in Section 2.2. However, the method applies straightforwardly, with the exception that on $\mathcal{D}_1$, we must model the joint conditional law $(Y_i, S_i) \mid X_i, W_i$ instead of the marginal conditional law $Y_i \mid X_i, W_i$. Of course, this is not hard: to do this, we can first fit (e.g.) a logistic regression to model $S_i \mid X_i, W_i$ and then fit another distributional regression to model $Y_i \mid X_i, S_i, W_i$, as in Section 2.4.

Although $\theta(P)$ is not an expectation over P, it can be reduced to a linear problem. In particular, note

$$\theta(P) = \frac{\mathbb{E}_P[(Y_i(1) - Y_i(0))\mathbb{I}(S_i(1) = S_i(0) = 1)]}{P(S_i(1) = S_i(0) = 1)}.$$

To analyze this, there are two cases. First, analysts often make assumptions which ensure that the denominator is identifiable (Lee, 2009; Semenova, 2021). In this case, we can first apply the standard dual bound methodology to the numerator, which is linear in P. Then, on the second fold $\mathcal{D}_2$, we also estimate the (identifiable) denominator. Finally, we combine estimates for the numerator and denominator using the bivariate delta method, as in Example 2 (see Appendix A.2 for an explicit calculation).

Yet even when the denominator is unidentifiable, $\theta(P)$ is still *quasilinear* in P. This means that $\theta(P) \leq c$ if and only if

$$\theta^{(c)}(P) \coloneqq \mathbb{E}_P[(Y_i(1) - Y_i(0))\mathbb{I}(S_i(1) = S_i(0) = 1)] - cP(S_i(1) = S_i(0) = 1) \leq 0.$$

Since the estimand $\theta^{(c)}(P)$ is linear in P, for any $c \in \mathbb{R}$, we can compute a lower confidence bound $\hat{\theta}^{(c)}_{\text{LCB}}$ for $\theta^{(c)}_L$, where $\theta^{(c)}_L$ is the lower bound on $\theta^{(c)}(P)$. Then, a valid lower confidence bound on θ_L is defined as

$$\hat{\theta}_L = \min\{c : \hat{\theta}^{(c)}_{\text{LCB}} \leq 0\}. \tag{18}$$

In practice, we can identify the minimum c in Eq. (18) using a grid search or binary search. This procedure is computationally tractable, although it is more expensive than the case where $\theta(P)$ is an expectation.

The ideas in Example 3 apply to any quasilinear function of P. Two examples are given below.

Example 5 (Conditional treatment effects)**.** Suppose $\theta(P) = \mathbb{E}_P[Y_i(1) - Y_i(0) \mid B]$, where B is some event which has strictly positive probability under any $P \in \mathcal{P}$. Then $\theta(P)$ is quasilinear in P, and we can compute valid dual bounds as in Example 3. One important special case is the subgroup treatment effect $\mathbb{E}[Y_i(1) - Y_i(0) \mid Y_i(0) \leq c]$ defined by Kaji and Cao (2023), where $c \in \mathbb{R}$ is a constant. If (e.g.) $Y(1), Y(0)$ measure income, Kaji and Cao (2023) interpreted this estimand as a treatment effect for disadvantaged individuals, e.g., individuals whose income would be below a certain level without the treatment.

Example 6 (Quantiles of the ITE)**.** Suppose $\theta(P) = Q_\alpha(Y_i(1) - Y_i(0))$, where $Q_\alpha(\cdot)$ denotes the α-quantile function. Then $\theta(P)$ is quasilinear in P (Boyd and Vandenberghe, 2004).

3 Theory

3.1 Uniform validity

Our first main theoretical result is that in randomized experiments, $\hat{\theta}_{\text{LCB}}$ is a valid $1 - \alpha$ lower confidence bound on θ_L under arbitrary model misspecification.

Assumption 3.1. *We assume the propensity scores $\pi(X_i) \coloneqq \mathbb{P}(W_i = 1 \mid X_i)$ are known and bounded away from zero and one, and the potential outcomes $(Y_i(1), Y_i(0))$ are conditionally independent of the treatment W_i given the covariates X_i.*

Note that the following result allows for the analyst to use any method to estimate the optimal dual variables $\hat{\nu}$. It also places no restrictions on the relationship between the potential outcomes and X_i, although we do require the following moment condition on $\hat{\nu}$.

Assumption 3.2. *For $k \in \{0, 1\}$, we assume the fourth moment $\mathbb{E}_P[\hat{\nu}_k(Y(k), X)^4 \mid \mathcal{D}_1] \leq B < \infty$ is bounded conditional on $\mathcal{D}_1$ and the conditional variance of $S_i = \frac{\hat{\nu}(Y_i, X_i)W_i}{\pi(X_i)} + \frac{\hat{\nu}(Y_i, X_i)(1-W_i)}{1-\pi(X_i)}$ is bounded away from zero, i.e., $\text{Var}_P(S_i \mid \mathcal{D}_1) \geq \frac{1}{B}$.*

Assumption 3.2 is weak, since (e.g.) in practice one could always "clip" $\hat{\nu}$ below some large value to ensure its moments exist. It can also be substantially relaxed at the cost of a more technical statement (see Appendix D.1, Remark 17). All we need is for the moments of S_i to be sufficiently regular such that we can apply a univariate central limit to $\{S_i\}_{i \in \mathcal{D}_2}$ conditional on $\mathcal{D}_1$.

Theorem 3.1. *Assume Assumption 3.1. For any $B \geq 0$, let $\mathcal{P}_B \subset \mathcal{P}$ denote the set of all laws $P \in \mathcal{P}$ such that $\hat{\nu}$ satisfies Assumption 3.2 under P. Then*

$$\liminf_{n \to \infty} \inf_{P \in \mathcal{P}_B} \mathbb{P}(\hat{\theta}_{\text{LCB}} \leq \theta_L) \geq 1 - \alpha.$$

Proof sketch. Note that if $\tilde{\theta}_L = g(\hat{\nu})$ as in Definition 1,

$$\theta_L - \hat{\theta}_{\text{LCB}} = \underbrace{\theta_L - \tilde{\theta}_L}_{\text{Term A}} + \underbrace{\tilde{\theta}_L - \hat{\theta}_{\text{LCB}}}_{\text{Term B}}.$$

Term A is positive deterministically by weak duality. Term B is positive with probability equal to $1 - \alpha$ asymptotically by standard results on IPW estimators under the conditions outlined in the theorem; in particular, it follows from the univariate Lyapunov CLT. □

Of course, by multiplying $\theta(P)$ by negative one, these theorems prove that we can get a $1-\alpha$ upper confidence bound $\hat{\theta}_{\text{UCB}}$ on the sharp upper bound θ_U. These bounds can be combined to cover either the partially identified set or the parameter $\theta(P^\star)$ (Imbens and Manski, 2004).

Theorem 3.1 has two key ingredients—(i) weak duality plus (ii) the fact that $\tilde{\theta}_L$ has a representation as a marginal moment, which allows us to apply the CLT. As discussed in Section 2.3, these properties also allow us to use the multiplier bootstrap to select a "good" choice of $\hat{\nu}$. In particular, the multiplier bootstrap is asymptotically valid as long as the central moments of the IPW summands $S_i^{(k)}$ do not grow too quickly with n and K, as stated formally below. The following assumption generally allows K to grow exponentially with a power of n, meaning that we can select from many different models without sacrificing validity.

Assumption 3.3 (Chernozhukov et al. (2018b))**.** *For K estimates $\hat{\nu}^{(1)}, \ldots, \hat{\nu}^{(K)}$ of $\nu^\star$, for $i \in \mathcal{D}_2$, define the IPW summands*

$$S_i^{(k)} \coloneqq \frac{\hat{\nu}_1^{(k)}(Y_i, X_i)W_i}{\pi(X_i)} + \frac{\hat{\nu}_0^{(k)}(Y_i, X_i)(1-W_i)}{1-\pi(X_i)} \text{ and } Z_{ik} = S_i^{(k)} - \mathbb{E}[S_i^{(k)} \mid \mathcal{D}_1].$$

We assume there exists $\epsilon \in (0, 1/4), c > 0$ such that

$$B_n \coloneqq \max_{k\in[K]} \left((\mathbb{E}[|Z_{ik}|^4 \mid \mathcal{D}_1)^{1/2} \vee (\mathbb{E}|Z_{ik}|^3 \mid \mathcal{D}_1) \right) + \mathbb{E}\left[\max_{k\in[K]} |Z_{ik}|^4 \mid \mathcal{D}_1 \right]^{1/4} \leq c \frac{n^{1/4-\epsilon}}{\log(Kn)^{7/4}}.$$

This assumption is quite weak and is standard in the literature. Note it is trivially satisfied if K and the moments of Z_{ik} do not grow with n.

Corollary 3.1. *Suppose the analyst computes K estimates $\hat{\nu}^{(1)}, \ldots, \hat{\nu}^{(K)}$ of $\nu^\star$ on $\mathcal{D}_1$ and uses the multiplier bootstrap to compute a lower bound $\hat{\theta}_{\text{LCB}}^{\text{MB}}$ as defined in Def. 2. Fix $c > 0, \epsilon \in (0, 1/4)$ and let $\mathcal{P}_{c,\epsilon}$ denote the set of laws $P \in \mathcal{P}$ such that Assumption 3.3 holds. Then under Assumption 3.1,*

$$\liminf_{n\to\infty} \inf_{P\in\mathcal{P}_{c,\epsilon}} \mathbb{P}(\hat{\theta}_{\text{LCB}}^{\text{MB}} \leq \theta_L) \geq 1-\alpha.$$

Each of these results relies on the weak duality result that $\tilde{\theta}_L \leq \theta_L$ holds deterministically. Although this ensures that $\hat{\theta}_{\text{LCB}}$ is a valid lower confidence bound, one might worry that it will make inference too conservative. We investigate this question in the next section.

3.2 Tightness

3.2.1 General analysis

We now give high-level conditions under which $\hat{\theta}_{\text{LCB}}$ converges to θ_L at oracle rates. The main intuition follows from the decomposition

$$\theta_L - \hat{\theta}_{\text{LCB}} = \underbrace{\theta_L - \tilde{\theta}_L}_{\text{first-stage bias}} + \underbrace{\tilde{\theta}_L - \hat{\theta}_{\text{LCB}}}_{\text{variance from the CLT}}.$$

The univariate CLT suggests that the second term is asymptotically exact. Thus, the main question is how large the first-stage bias is.

The following theorem tells us that the first stage bias is bounded by the product of the errors in estimating $(P^\star_{Y(0)|X}, P^\star_{Y(1)|X})$ and $\nu^\star$. Thus, if the product of the errors decays at an $o(n^{-1/2})$ rate, the first stage bias will be negligible compared to the variance from the univariate CLT. As notation, let $p_0^\star(y_0 \mid x), p_1^\star(y_1 \mid x)$ denote the conditional densities of $Y(0) \mid X$ and $Y(1) \mid X$ with respect to some base measure ψ_x on $\mathcal{Y}$ [3]; similarly, let $\hat{p}_1(y_1 \mid x), \hat{p}_0(y_0 \mid x)$ denote the estimated densities under $\hat{P}_{Y(0)|X}, \hat{P}_{Y(1)|X}$.

For each $x \in \mathcal{X}$, we define $\text{error}_P(x)$ to be the ℓ_2 distance between $(p_0^\star, p_1^\star)$ and $(\hat{p}_0, \hat{p}_1)$:

$$\text{error}_P(x) \coloneqq \left(\sum_{k\in\{0,1\}} \int_{y\in\mathcal{Y}} (p_k^\star(y \mid x) - \hat{p}_k(y \mid x))^2 \psi_x(dy) \right)^{1/2}. \tag{19}$$

[3] We choose ψ_x to be the Lebesgue measure for continuous potential outcomes and the counting measure for discrete potential outcomes.

Similarly, we define $\text{error}_\nu(x)$ to be the corresponding ℓ_2 distance between $\hat{\nu}$ and $\nu^\star$:

$$\text{error}_\nu(x) \coloneqq \left(\sum_{k\in\{0,1\}} \int_{y\in\mathcal{Y}} (\nu_k^\star(y,x) - \hat{\nu}_k(y,x))^2 \psi_x(dy) \right)^{1/2}. \tag{20}$$

Theorem 3.2. *Suppose strong duality holds, i.e., $g(\nu^\star) = \theta_L$. Then the first stage bias is bounded by the product of the errors in estimating the laws of $Y(k) \mid X, k \in \{0,1\}$ and the error in estimating $\nu^\star$. Formally,*

$$0 \le \theta_L - \tilde{\theta}_L \le \mathbb{E}\left[\text{error}_P(X) \cdot \text{error}_\nu(X)\right]. \tag{21}$$

Overall, Theorem 3.2 gives intuition that if strong duality holds, the first stage bias should decay at a faster rate than $\mathbb{E}_{X\sim P_X^\star}[\text{error}_P(x)]$, which represents the error in estimating the outcome model. Intuitively, this is because if $\hat{P}_{Y(0)|X}, \hat{P}_{Y(1)|X}$ is close to $P^\star_{Y(0)|X}, P^\star_{Y(1)|X}$, then $\hat{\nu}$ should be close to $\nu^\star$, since $\hat{\nu}$ maximizes the empirical dual based on $\hat{P}_{Y(0)|X}, \hat{P}_{Y(1)|X}$, and $\nu^\star$ solves the population dual based on $P^\star_{Y(0)|X}, P^\star_{Y(1)|X}$.

We emphasize that as long as strong duality holds, Theorem 3.2 makes no assumptions whatsoever about the form of $\theta(P)$, the dimension of the covariates X, or the model class $\mathcal{P}$—furthermore, it is a finite-sample result with no "hidden" constants. Thus, Theorem 3.2 is very useful in analyzing the tightness of dual bounds in a wide variety of asymptotic regimes, including high-dimensional regimes where the dimension of X grows with n.

3.2.2 Refined theory for discrete potential outcomes

Previously, we used Theorem 3.2 to argue that the first-stage bias of dual bounds decays faster than the estimation error of the outcome model $\text{error}_P(X)$. Now, we formalize this intuition in the case where Y has finite support and $\hat{\nu}$ are chosen as the dual variables corresponding to $(\hat{P}_{Y(0)|X}, \hat{P}_{Y(1)|X})$. In particular, using a technical tool called *Hoffman constants* (Hoffman, 1952), Lemma 3.3 now shows that for each $x \in \mathcal{X}$, the error in estimating the dual variables decays linearly in $\text{error}_P(x)$.

Lemma 3.3. *Suppose $\mathcal{Y}$ is finite and consider estimated dual variables $\hat{\nu}$ as defined in Eq. (12). There exist a collection of finite deterministic Lipschitz constants $\{H(x) : x \in \mathcal{X}\}$ depending only on $P^\star$, $\mathcal{P}$ and f such that for all $x \in \mathcal{X}$, there exist optimal dual variables $\nu^\star$ such that the following holds deterministically:*

$$\text{error}_\nu(x)^2 \coloneqq \sum_{k\in\{0,1\}} \sum_{y\in\mathcal{Y}} (\hat{\nu}_k(y,x) - \nu^\star(y,x))^2 \le H(x) \cdot \text{error}_P(x)^2.$$

Note that Lemma 3.3 allows for settings where the optimal dual variables $\nu^\star$ are not unique. Nonetheless, there always exists *some* choice of $\nu^\star \in \arg\max_\nu g(\nu)$ such that Lemma 3.3 holds.

Combining Theorem 3.2 and Lemma 3.3 establishes that if the error in estimating the outcome model, $\text{error}_P(x)$, decays at $o(n^{-1/4})$ rates, then $\hat{\theta}_L$ and $\hat{\theta}_{\text{LCB}}$ are equivalent to "oracle" estimators which have perfect knowledge of the outcome model and use the optimal dual variables $\nu^\star$ in place of $\hat{\nu}$. More precisely, although $\nu^\star$ may not be unique, there exists some sequence of optimal dual variables $\nu_n^\star \in \arg\max_\nu g(\nu)$ such that $\hat{\theta}_L, \hat{\theta}_{\text{LCB}}$ are asymptotically equivalent to the oracle estimators based on $\nu_n^\star$. As notation, for any sequence of random variables $\{R_n\}_{n\in\mathbb{N}}$ and any sequence $\{a_n\}_{n\in\mathbb{N}}$, we say $R_n = o_{L_k}(a_n)$ if and only $\mathbb{E}[(R_n/a_n)^k] \to 0$ as $n \to \infty$.

Theorem 3.4. *Suppose Y has finite support $\mathcal{Y}$ and $\mathbb{E}[|H(X)|^2] < \infty$. Furthermore, assume that $\text{error}_P(X) = o_{L_4}(n^{-1/4})$ as $n \to \infty$, where X denotes a fresh sample of covariates. Then there exist a sequence of optimal dual variables $\nu_n^\star \in \arg\max_{\nu\in\mathcal{V}} g(\nu)$ such that if $\hat{\theta}_L^\star$, $\hat{\theta}_{\text{LCB}}^\star$ denote the oracle estimators which use $\nu_n^\star$ in place of $\hat{\nu}$,*

$$\sqrt{n}(\hat{\theta}_L - \hat{\theta}_L^\star) \xrightarrow{\text{p}} 0 \text{ and } \sqrt{n}(\hat{\theta}_{\text{LCB}} - \hat{\theta}_{\text{LCB}}^\star) \xrightarrow{\text{p}} 0. \tag{22}$$

Theorem 3.4 shows that as long as one can estimate the conditional laws of the potential outcomes at semi-parametric rates, then the dual bounds $\hat{\theta}_L, \hat{\theta}_{\text{LCB}}$ asymptotically perform as well as an oracle estimator with perfect knowledge of the laws of $(Y(1), X)$ and $(Y(0), X)$. By Theorem 3.1, the oracle estimator converges to the sharp lower bound $\theta_L = \max_{\nu\in\mathcal{V}} g(\nu)$ with the parametric rate. Thus, under the assumptions of Theorem 3.4, $\hat{\theta}_L$ is root-n consistent for the sharp lower bound θ_L with the same asymptotic variance with the oracle estimator.

Remark 3 (Discussion of Assumptions)**.** Theorem 3.4 makes two main assumptions besides the hypothesis that $\text{error}_P(X) = o_{L_4}(n^{-1/4})$.

1. A restrictive assumption is that Y has a finite support. One could try to approximate any continuous distribution by allowing $|\mathcal{Y}|$ to grow with n, but we leave this to future work. Nonetheless, the intuition of Theorem 3.2 suggests that a result similar to Theorem 3.4 likely holds in the continuous case.

2. Theorem 3.4 also requires that $H(X)$ has at least two moments. Since $H(X)$ is provably a finite-valued random variable, we do not think this assumption is too restrictive, especially since the law of $H(X)$ only depends on population quantities; additionally, we show in Appendix D.3 that the moments of $H(X)$ generally do not grow with the dimension of X. Furthermore, we can show that if a certain "general position" condition holds on the conditional probability mass functions of $Y(k) \mid X$, then this moment condition is satisfied. However, this analysis is rather technical, so we defer it to Appendix D.5.

Remark 4 (Comparison to Semenova (2023))**.** Theorem 3.4 has a similar flavor to Theorem 3.1 proved in Semenova (2023). However, we use a completely different proof technique, which yields a complementary result that is stronger in some ways. For instance, Semenova (2023) requires that $\sup_{x\in\mathcal{X}} \text{error}_P(x) = o_p(n^{-1/4})$. This may not be realistic when $\mathcal{X}$ is a large continuous set. We only require the weaker condition that $\text{error}_P(X) = o_{L_4}(n^{-1/4})$. Furthermore, Semenova (2023) does not apply to $\hat{\nu}$, but rather applies to a different estimator for which the computation time is potentially exponential in $|\mathcal{Y}|$. Thus, a major benefit of Theorem 3.4 is that one can compute $\hat{\nu}$ efficiently.

Before concluding this section, we note an interesting theoretical curiosity: in low dimensions, it is possible to dramatically strengthen Theorem 3.4. Indeed, in Appendix D.3, we show that under technical (and strong) regularity conditions, if $\sup_{x\in\mathcal{X}} \text{error}_P(x)$ converges to zero at any rate, then the estimator $\hat{\theta}_L$ will deterministically equal the oracle estimator $\hat{\theta}_L^\star$ with probability one. That said, we emphasize that this result requires strong assumptions and is very asymptotic in nature, and in most practical applications, we do not expect that $\hat{\theta}_L = \hat{\theta}_L^\star$ will hold. Thus, Theorem 3.4 is more informative in practice.

3.3 Cross fitting

In Section 3.2, we saw that if the outcome model can be estimated at $o(n^{-1/4})$ rates, then $\hat{\theta}_L$ is equivalent to an oracle estimator $\hat{\theta}_L^\star$ which has full knowledge of the outcome model. However, both $\hat{\theta}_L$ and $\hat{\theta}_L^\star$ use only half the data to compute their respective IPW estimators. A stronger oracle is the estimator

$$\hat{\theta}_L^{\star\star} := \frac{1}{n}\sum_{i=1}^{n} \frac{\nu_1^\star(Y_i, X_i)W_i}{\pi(X_i)} + \frac{\nu_0^\star(Y_i, X_i)(1-W_i)}{1-\pi(X_i)} \tag{23}$$

which uses the full dataset to compute the oracle IPW estimator; as a result, its variance should be twice as small. A corresponding oracle lower confidence bound $\hat{\theta}_{\text{LCB}}^{\star\star}$ can be computed by subtracting off the sample standard deviation of the summands in (23) times $\Phi^{-1}(1-\alpha)n^{-1/2}$. In this section, we show that employing cross-fitting can recover this factor of two. Furthermore, cross-fitting is still valid under most forms of misspecification.

To define notation, let $\hat{\theta}_L^{\text{swap}}$ denote the same estimator as θ_L but with the roles of $\mathcal{D}_1$ and $\mathcal{D}_2$ swapped. The cross-fit estimator is then

$$\hat{\theta}_L^{\text{crossfit}} := \frac{\hat{\theta}_L + \hat{\theta}_L^{\text{swap}}}{2}.$$

A cross-fit lower confidence bound can be computed as follows. Let $\hat{\nu}$ and $\hat{\nu}^{\text{swap}}$ denote the estimated dual variables from $\mathcal{D}_1$ and $\mathcal{D}_2$, respectively. For ease of exposition, we assume n is even and $|\mathcal{D}_1| = |\mathcal{D}_2| = n/2$. [4] Let $S_i = \frac{\hat{\nu}_1^{\text{swap}}(Y_i,X_i)W_i}{\pi(X_i)} + \frac{\hat{\nu}_0^{\text{swap}}(Y_i,X_i)(1-W_i)}{1-\pi(X_i)}$ if $i \in \mathcal{D}_1$. If $i \in \mathcal{D}_2$, let S_i be defined analogously but with $\hat{\nu}^{\text{swap}}$ replaced with $\hat{\nu}$. Then if $\hat{\sigma}_s^{\text{crossfit}}$ is the empirical standard deviation of $\{S_i\}_{i=1}^n$, the cross-fit lower confidence bound is

$$\hat{\theta}_{\text{LCB}}^{\text{crossfit}} = \hat{\theta}_L^{\text{crossfit}} - \Phi^{-1}(1-\alpha)\frac{\hat{\sigma}_s^{\text{crossfit}}}{\sqrt{n}}. \tag{24}$$

We now prove that under the same conditions as in the previous section, $\hat{\theta}_L^{\text{crossfit}}$ and $\hat{\theta}_{\text{LCB}}^{\text{crossfit}}$ are equivalent to the oracle versions. This is relatively simple to prove; for example, the result for $\hat{\theta}_L^{\text{crossfit}}$ follows immediately by applying Theorem 3.4 to both $\hat{\theta}_L$ and $\hat{\theta}_L^{\text{swap}}$.

[4] The results in this section can be easily extended to M-fold cross-fitting for $M > 2$.

Proposition 3.1. *Assume the conditions of Theorem 3.4. Then*

$$\sqrt{n}(\hat{\theta}_L^{\text{crossfit}} - \hat{\theta}_L^{\star\star}) \xrightarrow{\text{p}} 0 \text{ and } \sqrt{n}(\hat{\theta}_{\text{LCB}}^{\text{crossfit}} - \hat{\theta}_{\text{LCB}}^{\star\star}) \xrightarrow{\text{p}} 0. \tag{25}$$

Proposition 3.1 tells us that, for discrete potential outcomes, if $\text{error}_P(X) = o_{L_4}(n^{-1/4})$ for $k \in \{0,1\}$, then $\hat{\theta}_L^{\text{crossfit}}$ is equivalent to the oracle; an immediate consequence is that $\hat{\theta}_{\text{LCB}}^{\text{crossfit}}$ is a valid asymptotic $1-\alpha$ lower bound on θ_L.

Now we turn to the general case where $\hat{\nu}$ are potentially inconsistent. Due to the dependence introduced by cross-fitting, we need more regularity conditions to show an analogue of Theorem 3.1. Interestingly, we show that $\hat{\theta}_{\text{LCB}}^{\text{crossfit}}$ is valid under two separate and non-nested conditions. Note that the following theorem does not require any of the assumptions of the previous theory, e.g. it does not require Y to be discrete and it does not require $\hat{\nu}$ to be computed in the method suggested in Definition 1.

Proposition 3.2. *Assume that $\hat{\nu}^{\text{swap}}$ is computed using the same procedure as $\hat{\nu}$ (but applied to $\mathcal{D}_2$ instead of $\mathcal{D}_1$), so that Assumption 3.2 holds for $\hat{\nu}^{\text{swap}}$. Under Assumption 3.1,*

$$\liminf_{n\to\infty} \mathbb{P}(\hat{\theta}_{\text{LCB}}^{\text{crossfit}} \le \theta_L) \ge 1-\alpha,$$

if one of the following holds:

1. *Condition 1: There exist arbitrary deterministic functions $\nu_k^\dagger : \mathcal{Y} \times \mathcal{X} \to \mathbb{R}$ satisfying the moment conditions in Assumption 3.2 such that $\mathbb{E}\left[\left(\hat{\nu}_k(Y(k), X) - \nu_k^\dagger(Y(k), X)\right)^2\right] \to 0$ holds at any rate for $k \in \{0,1\}$. Note that we allow $\{\nu_k^\dagger\}_{k\in\{0,1\}}$ to change with n.*
2. *Condition 2: The outcome model is sufficiently misspecified such that the first-stage bias is larger than $n^{-1/2}$, i.e., $n^{-1/2}(\theta_L - \tilde{\theta}_L) \xrightarrow{\text{p}} \infty$.*

The first condition of Proposition 3.2 is that as long as the estimated dual functions $\hat{\nu}_0, \hat{\nu}_1$ are asymptotically deterministic, though the limits may differ from $(\nu_0^\star, \nu_1^\star)$, $\hat{\theta}_{\text{LCB}}^{\text{crossfit}}$ is a valid lower confidence bound. Similar conditions on estimated nuisance parameters have been studied in other contexts (Chernozhukov et al., 2020; Arkhangelsky et al., 2021). A strength of this result is that it allows $\hat{\nu}_0, \hat{\nu}_1$ to converge at arbitrarily slow rates. Indeed, the proof technique for this result is based on a novel argument leveraging weak duality; it is not necessarily true that under Condition 1, $\hat{\theta}_{\text{LCB}}^{\text{crossfit}}$ is equivalent to an "oracle" confidence bound of any form. The second condition suggests that even if this is not true and the fluctuations of $\hat{\nu}_0, \hat{\nu}_1$ do not vanish asymptotically, cross-fitting can be valid if the first-stage bias is sufficiently large, making $\hat{\theta}_{\text{LCB}}^{\text{crossfit}}$ conservative but valid.

Except in pathological examples, we expect the second condition to hold whenever the first condition does not. Intuitively, if $\hat{\nu}$ has non-vanishing fluctuations, this suggests that $\hat{\nu}$ is not consistently estimating $\nu^\star$, in which case we should expect $O(1)$ conservative bias, satisfying Condition 2. Our simulations in Section 5.4 confirm this intuition. Thus, in practice, we recommend using cross-fitting to reduce the variance of the final estimator.

Remark 5. Under the conditions of Proposition 3.2, one can use cross-fitting in combination with a multiplier-bootstrap-like procedure to perform model selection as long as one chooses among a finite number of (fit) outcome models. We present this result in Appendix A.2 for brevity.

3.4 Dual bounds for observational studies

So far, our theory has assumed that the propensity scores are known. However, when $\pi(X_i)$ is unknown, we can replace the IPW estimator with an AIPW estimator to increase robustness. In particular, for the estimated dual variables, $\hat{\nu}_0, \hat{\nu}_1 : \mathcal{Y} \times \mathcal{X} \to \mathbb{R}$, define their conditional means as

$$c_0(x) \coloneqq \mathbb{E}_{P^\star_{Y(0)|X=x}}[\hat{\nu}_0(Y(0), x)] \text{ and } c_1(x) \coloneqq \mathbb{E}_{P^\star_{Y(1)|X=x}}[\hat{\nu}_1(Y(1), x)]$$

so, e.g., $c_0(X_i)$ is the conditional mean of $\hat{\nu}_0(Y(0), X_i)$ given X_i and $\mathcal{D}_1$. Also, let $\hat{c}_0(x), \hat{c}_1(x)$ denote estimators of $c_0(x), c_1(x)$ fit on $\mathcal{D}_1$; for example, one can automatically compute $\hat{c}_0(x), \hat{c}_1(x)$ by plugging in $\hat{P}_{Y(0)|X}, \hat{P}_{Y(1)|X}$. Lastly, for any $i \in \mathcal{D}_2$, define the AIPW summand

$$S_i \coloneqq W_i \frac{\hat{\nu}_1(Y_i, X_i) - \hat{c}_1(X_i)}{\hat{\pi}(X_i)} + (1-W_i)\frac{\hat{\nu}_0(Y_i, X_i) - \hat{c}_0(X_i)}{1-\hat{\pi}(X_i)} + \hat{c}_1(X_i) + \hat{c}_0(X_i), \tag{26}$$

where $\hat{\pi}$ are propensity scores estimated on $\mathcal{D}_1$. Then, if $\hat{\sigma}_s^{\text{aug}}$ is the sample standard deviation of $\{S_i\}_{i\in\mathcal{D}_2}$ on $\mathcal{D}_2$, the "augmented" version of $\hat{\theta}_{\text{LCB}}$ is

$$\hat{\theta}_{\text{LCB}}^{\text{aug}} \coloneqq \frac{1}{n_2}\sum_{i\in\mathcal{D}_2} S_i - \Phi^{-1}(1-\alpha)\frac{\hat{\sigma}_s^{\text{aug}}}{\sqrt{n_2}}. \tag{27}$$

We can now prove a validity result for $\hat{\theta}_{\text{LCB}}^{\text{aug}}$. There are two cases. In the first case, we assume that the product of estimation errors for the outcome model and propensity scores decays faster than $o(1/n)$, in which case $\hat{\theta}_{\text{LCB}}^{\text{aug}}$ will be a valid lower confidence bound for $\tilde{\theta}_L$ based on standard results for the AIPW estimator (Wager, 2020). However, even outside this standard regime, $\hat{\theta}_{\text{LCB}}^{\text{aug}}$ may still be valid. In the second case, we assume the outcome model is sufficiently misspecified such that the first stage bias $\theta_L - \tilde{\theta}_L$ dominates either the error in estimating π or the error in estimating c. In this situation, the fluctuations of $\hat{\theta}_{\text{LCB}}^{\text{aug}}$ around $\tilde{\theta}_L$ are of smaller order than the first-stage bias. Intuitively, this second case explains a phenomenon we have observed empirically, which is that even when one cannot consistently estimate the propensity scores or outcome model, the first-stage bias $\tilde{\theta}_L - \theta_L$ often (but not always) makes $\hat{\theta}_{\text{LCB}}^{\text{aug}}$ conservative.

Theorem 3.5. *Suppose Assumption 3.1 holds except that the propensity scores are not known. For $k \in \{0,1\}$, assume the fourth moments $\mathbb{E}[|\hat{\nu}_k(Y(k),X)|^4 \mid \mathcal{D}_1] \le B < \infty$ and $\mathbb{E}[|\hat{c}_k(X)|^4 \mid \mathcal{D}_1] \le B < \infty$ are uniformly bounded. Finally, assume that the estimated propensity scores $\hat{\pi}(X_i)$ are uniformly bounded away from zero and one.*

Let $\text{error}_n(\hat{\pi}) \coloneqq \mathbb{E}[(\hat{\pi}(X)-\pi(X))^2 \mid \mathcal{D}_1]^{1/2}$ denote the ℓ_2 error in estimating the propensity scores and let $\text{error}_n(\hat{c}) = \max_{k\in\{0,1\}} \mathbb{E}[(\hat{c}_k(X)-c_k(X))^2 \mid \mathcal{D}_1]^{1/2}$ denote the ℓ_2 error in estimating the conditional mean of $\hat{\nu}$, where X is an independent draw from the law of X_i. Consider the two conditions below:

- *Condition 1:* $\text{error}_n(\hat{\pi}) = o_{L_2}(1)$, $\text{error}_n(\hat{c}) = o_{L_2}(1)$, *and the "risk-decay" condition holds:*

$$\mathbb{E}[\text{error}_n(\hat{\pi})^2]\mathbb{E}[\text{error}_n(\hat{c})^2] = o(1/n).$$

  Furthermore, if $\tilde{S}_i = W_i\frac{\hat{\nu}_1(Y_i,X_i)-c_1(X_i)}{\pi(X_i)} + (1-W_i)\frac{\hat{\nu}_0(Y_i,X_i)-c_0(X_i)}{1-\pi(X_i)} + c_1(X_i) + c_0(X_i)$, we assume $\text{Var}(\tilde{S}_i \mid \mathcal{D}_1) \ge \frac{1}{B}$ is bounded away from zero.

- *Condition 2: the outcome model is sufficiently misspecified such that the first-stage bias $\tilde{\theta}_L - \theta_L$ dominates either* $\text{error}_n(\hat{\pi})$ *or* $\text{error}_n(\hat{c})$. *More precisely, assume*

$$\frac{\min(\text{error}_n(\hat{c}), \text{error}_n(\hat{\pi}))}{\tilde{\theta}_L - \theta_L} \xrightarrow{\text{p}} 0.$$

If either Condition 1 or Condition 2 holds, then $\hat{\theta}_{\text{LCB}}^{\text{aug}}$ is asymptotically valid:

$$\liminf_{n\to\infty} \mathbb{P}(\hat{\theta}_{\text{LCB}}^{\text{aug}} \le \theta_L) \ge 1-\alpha.$$

See Appendix D.1 for a proof.

Remark 6. In observational studies, the multiplier bootstrap method for model selection from Section 2.3 is not appropriate because the validity of the final bounds may depend on the accuracy of the outcome model. For example, the multiplier bootstrap might select a highly inaccurate outcome model that yields (misleadingly) tight bounds. Thus, in observational studies, we recommend that the analyst perform cross-validation on $\mathcal{D}_1$ to select the best-performing outcome model.

4 Computation

4.1 General strategy and ensuring validity

In this section, we discuss how to compute the dual bounds in Definition 1. Computation is straightforward except for two questions:

- Dual bounds will yield valid results for any estimated dual variables as long as $(\hat{\nu}_0, \hat{\nu}_1) \in \mathcal{V}$ are dual-feasible. However, it is not obvious how to ensure that dual-feasibility holds.

- Our recommended approach to estimating the dual variables requires solving the optimization problem

$$\hat{\nu}_{0,x}, \hat{\nu}_{1,x} = \underset{(\nu_{0,x},\nu_{1,x})\in\mathcal{V}_x}{\arg\max} \mathbb{E}_{\hat{P}_{Y(0)|X}}[\nu_{0,x}(Y(0)) \mid X = x] + \mathbb{E}_{\hat{P}_{Y(1)|X}}[\nu_{1,x}(Y(1)) \mid X = x]. \tag{28}$$

  If Y is continuous, this is an infinite-dimensional program, so it is unclear how to solve it.

We now outline a general strategy to answer these questions based on two key observations. Note that for simplicity, in this section, we assume the response Y is real-valued.

Observation 1: the problem separates in $\mathcal{X}$. Theorem 2.1 makes clear that to compute $\hat{\nu} = \arg\max_{\nu\in\mathcal{V}} g(\hat{\nu})$, it suffices to repeatedly solve the problem conditional on x. I.e., it suffices to solve for $\hat{\nu}_{0,x}, \hat{\nu}_{1,x}$ as per Eq. (28), and the solutions to these problems are independent in the sense that the value of $\hat{\nu}_{0,x}, \hat{\nu}_{1,x}$ does not affect the value of $\hat{\nu}_{0,x'}, \hat{\nu}_{1,x'}$ for some $x' \neq x$. Similarly, by definition of $\mathcal{V}$ we have that $\hat{\nu} \in \mathcal{V}$ is valid if and only if $\hat{\nu}_{0,x}, \hat{\nu}_{1,x} \in \mathcal{V}_x$ for all $x \in \mathcal{X}$.

Observation 2: Only compute what we need. To apply dual bounds, we only need to compute the values of $\hat{\nu}_0, \hat{\nu}_1$ on the second fold of data $\mathcal{D}_2$. I.e., we need only compute $\{\hat{\nu}_{0,x}, \hat{\nu}_{1,x}\}_{x\in\{X_i:i\in\mathcal{D}_2\}}$ to compute the IPW estimator $\hat{\theta}_L$ and lower confidence bound $\hat{\theta}_{\text{LCB}}$. Note that there is no problem of "double dipping" or "selective inference" because the final estimator based only on $\{\hat{\nu}_{0,x}, \hat{\nu}_{1,x}\}_{x\in\{X_i:i\in\mathcal{D}_2\}}$ is identical to the final estimator based on the full functions $\hat{\nu}_0, \hat{\nu}_1$.

These observations have two implications.

Implication 1: ensuring validity. Given *any* initial estimators $\hat{\nu}_0^{\text{init}}, \hat{\nu}_1^{\text{init}}$ which may or may not be dual-feasible, we can convert $\hat{\nu}_0^{\text{init}}, \hat{\nu}_1^{\text{init}}$ into dual-feasible estimators as follows:

- For $x \in \mathcal{X}$, we define c_x to be half of the maximum violation of the conditional feasibility constraint:

$$2c_x \coloneqq \max_{y_1,y_0\in\mathcal{Y}} \hat{\nu}_{0,x}^{\text{init}}(y_0) + \hat{\nu}_{1,x}^{\text{init}}(y_1) - f(y_1, y_0, x).$$

  Alternatively, for any estimated $\hat{\lambda}_{x,1}, \ldots, \hat{\lambda}_{x,L} \geq 0$, we could also define

$$2c_x \coloneqq \max_{y_1,y_0\in\mathcal{Y}} \hat{\nu}_{0,x}^{\text{init}}(y_0) + \hat{\nu}_{1,x}^{\text{init}}(y_1) - \sum_{\ell=1}^{L} \hat{\lambda}_{x,\ell} w_{x,\ell} - f(y_1, y_0, x).$$

  However, for simplicity, we emphasize that choosing $\hat{\lambda}_{x,1} = \cdots = \hat{\lambda}_{x,L} = 0$ is always a valid choice.
- Then we define the final estimators

$$\hat{\nu}_{0,x}(y_0) \coloneqq \hat{\nu}_{0,x}^{\text{init}}(y_0) - c_x \text{ and } \hat{\nu}_{1,x}(y_1) \coloneqq \hat{\nu}_{1,x}^{\text{init}}(y_1) - c_x \tag{29}$$

  which are guaranteed to be dual-feasible by definition of $\mathcal{V}$.

For each $x \in \mathcal{X}$, c_x can be computed using a two-dimensional grid search—crucially, because this grid search is low-dimensional, we can accurately compute c_x. Furthermore, Observation 2 implies that we only need to compute c_x for $\{X_i : i \in \mathcal{D}_2\}$. As a result, the steps above represent a generic algorithm to convert *any* estimates $\hat{\nu}_0^{\text{init}}, \hat{\nu}_1^{\text{init}}$ into valid dual estimators $\hat{\nu}_0, \hat{\nu}_1 \in \mathcal{V}$ via $O(n)$ grid searches.

Implication 2: a generic strategy for computing optimal dual variables. Similarly, to compute $\hat{\nu}_0, \hat{\nu}_1 = \arg\max_{\nu\in\mathcal{V}} g(\nu)$, we have the following general strategy:

- Step 1: Estimate $\hat{P}_{Y(0)|X}, \hat{P}_{Y(1)|X}$ on $\mathcal{D}_1$.
- Step 2: For $i \in \mathcal{D}_2$, solve the "conditional problem" Eq. (28) for $x = X_i$ and use the outputs $\hat{\nu}_{0,X_i}, \hat{\nu}_{1,X_i}$ to compute the inverse-propensity weighted summands in the definition of $\hat{\theta}_L$ and $\hat{\theta}_{\text{LCB}}$.

In other words, we need to only solve the conditional problem $O(n)$ times to compute the dual bounds. Although it is still not clear how to solve the conditional problem, we discuss how to do this in the next section. We emphasize that this is much easier than directly solving the unconditional problem $\hat{\nu} = \arg\max_{\nu\in\mathcal{V}} g(\nu)$, because in the unconditional problem, the optimization variables $\nu_0, \nu_1 : \mathcal{Y} \times \mathcal{X} \to \mathbb{R}$ are potentially high-dimensional functions. In contrast, the conditional problem (Eq. (28)) is still an infinite-dimensional problem, but the optimization variables $\nu_{0,x}, \nu_{1,x} : \mathcal{Y} \to \mathbb{R}$ are univariate functions. Thus, there is some hope that we can solve Eq. (28) a finite number of times.

Remark 7. We emphasize that no matter how poorly we solve Eq. (28), as long as we adjust our final dual variables using Eq. (29), we will get valid lower confidence bounds on θ_L.

4.2 Finding conditionally optimal dual variables

In Section 4.1, we reduced the problem of estimating the plug-in dual estimator $\hat{\nu} = \arg\max_\nu \hat{g}(\hat{\nu})$ to the problem of solving the conditional problem for $x \in \{X_i : i \in \mathcal{D}_2\}$:

$$\hat{\nu}_{0,x}, \hat{\nu}_{1,x} = \underset{(\nu_{0,x},\nu_{1,x})\in\mathcal{V}_x}{\arg\max} \mathbb{E}_{\hat{P}_{Y(0)|X}}[\nu_{0,x}(Y(0)) \mid X = x] + \mathbb{E}_{\hat{P}_{Y(1)|X}}[\nu_{1,x}(Y(1)) \mid X = x].$$

In this section, we give two methods to approximately solve this conditional problem and obtain initial estimates $\hat{\nu}^{\text{init}}_{0,x}, \hat{\nu}^{\text{init}}_{1,x}$. Then, we can use the two-dimensional grid search procedure outlined in Section 4.1 to convert these estimates into a final estimator $(\hat{\nu}_{0,x}, \hat{\nu}_{1,x}) \in \mathcal{V}_x$ which is guaranteed to be dual-feasible and provide a valid confidence bound on θ_L.

Method 1: Discretization. The first approach is to approximate $\hat{P}_{Y(k)|X=x}$ as a discrete distribution with support $\{y_{k,1,x}, \ldots, y_{k,n_{\text{vals}},x}\}$ and PMF $p_{k,1,x}, \ldots, p_{k,n_{\text{vals}},x} \in (0,1)$ so that

$$\hat{P}_{Y(0)|X=x} \approx \sum_{j=1}^{n_{\text{vals}}} p_{0,j,x}\delta_{y_{0,j,x}} \text{ and } \hat{P}_{Y(1)|X=x} \approx \sum_{i=1}^{n_{\text{vals}}} p_{1,i,x}\delta_{y_{1,i,x}},$$

where δ_z denotes the delta dirac measure on $z \in \mathbb{R}$. For example, we suggest taking $y_{k,j,x}$ as the $\frac{j}{n_{\text{vals}}+1}$th quantile of $\hat{P}_{Y(k)|X=x}$ and setting $p_{k,j,x} = \frac{1}{n_{\text{vals}}}$ for $k \in \{0,1\}, j \in \{1, \ldots, n_{\text{vals}}\}$. After making this approximation, the conditional optimization problem becomes a discrete linear program with $2n_{\text{vals}}$ variables and n^2_{vals} constraints:

$$\begin{aligned}
&\max \sum_{j=1}^{n_{\text{vals}}} p_{0,j,x}\nu_{0,x}(y_{0,j,x}) + \sum_{i=1}^{n_{\text{vals}}} p_{1,i,x}\nu_{1,x}(y_{1,i,x}) \\
&\text{s.t. } \nu_{0,x}(y_{0,j,x}) + \nu_{1,x}(y_{1,i,x}) - \sum_{\ell=1}^{L} \lambda_{x,\ell} w_{x,\ell}(y_{0,j,x}, y_{1,i,x}) \le f(y_{0,j,x}, y_{1,i,x}, x) \text{ for all } i,j \in [n_{\text{vals}}] \\
&\quad \lambda_{x,1}, \ldots, \lambda_{x,L} \ge 0,
\end{aligned}$$

where the optimization variables are $\{\nu_{0,x}(y_{0,j,x})\}_{j=1}^{n_{\text{vals}}}, \{\nu_{1,x}(y_{1,i,x})\}_{i=1}^{n_{\text{vals}}}$ and $\lambda_{x,1}, \ldots, \lambda_{x,L}$. We remind the reader that the functions $\{w_{x,\ell}\}_{\ell=1}^{L} = \mathcal{W}_x$ represent the a-priori conditional moment inequality constraints on $\mathcal{P}$, and in the default scenario where $\mathcal{P}$ is unconstrained, $\mathcal{W}_x$ is empty and all terms involving $w_{x,\ell}$ and $\lambda_{x,\ell}$ can be ignored.

This is a linear program with $2n_{\text{vals}} + L$ variables and $n^2_{\text{vals}} + L$ constraints, and it can be solved very efficiently using off-the-shelf LP solvers if (e.g.) $n_{\text{vals}} \le 100$. After solving this problem, we obtain initial values $\{\hat{\nu}^{\text{init}}_{0,x}(y_{0,j,x})\}_{j=1}^{n_{\text{vals}}}, \{\hat{\nu}^{\text{init}}_{1,x}(y_{1,i,x})\}_{i=1}^{n_{\text{vals}}}$ and we define the full functions $\hat{\nu}_{0,x}, \hat{\nu}_{1,x} : \mathbb{R} \to \mathbb{R}$ via (e.g.) linear interpolation. Then, as described in Section 4.1, we can use a two-dimensional grid search to obtain valid dual variables $\hat{\nu}_{0,x}, \hat{\nu}_{1,x}$.

Method 2: Approximation via basis functions. A second option is to choose a collection of basis functions $\phi_m : \mathcal{Y} \to \mathbb{R}$ for $m = 1, \ldots, M \in \mathbb{N}$ and approximate

$$\nu_{k,x}(y) \approx \sum_{m=1}^{M} \alpha_{k,m,x}\phi_m(y) \text{ for } \alpha_{k,m,x} \in \mathbb{R}, k \in \{0,1\}, m \in [M].$$

This reduces the problem to fitting the values of $\{\alpha_{k,m,x}\}_{m\in M, k\in\{0,1\}} \in \mathbb{R}^{2M}$, which is a concave problem with finitely many parameters. Of course, approximating the dual variables with a finite collection of basis functions introduces approximation errors, but we emphasize that our approach still yields *valid* bounds even if our initial estimates $\hat{\nu}^{\text{init}}$ based on the basis functions are arbitrarily poor. Furthermore, selecting a collection of universal basis functions (e.g., splines, Fourier expansion, Gaussian kernel) can ensure that the approximation errors are not too large.

However, fitting $\{\alpha_{k,m,x}\}_{k\in\{0,1\},m\in[M]}$ is still challenging because the conditional validity constraint $(\nu_{0,x}, \nu_{1,x}) \in \mathcal{V}_x$ is still infinite-dimensional. To overcome this, we use ideas from the optimal transport literature. Indeed, for this particular problem, we can eliminate the effect of the constraints by adding the maximum deviation from the constraints as the penalty in an objective. In particular, for the product measure

$\hat{P}_{\text{prod},x} \coloneqq \hat{P}_{Y(1)|X=x} \times \hat{P}_{Y(0)|X=x}$, consider the objective

$$
\begin{aligned}
&O(\nu_{0,x}, \nu_{1,x}, \{\lambda_{x,\ell}\}_{\ell=1}^{L}) \\
\coloneqq &\mathbb{E}_{\hat{P}_{\text{prod},x}}\Bigg[\nu_{0,x}(Y(0)) + \nu_{1,x}(Y(1)) - \underbrace{\max_{y_1,y_0 \in \mathcal{Y}} \nu_{0,x}(y_0) + \nu_{1,x}(y_1) - \sum_{\ell=1}^{L} \lambda_{x,\ell} w_{x,\ell}(y_0, y_1) - f(y_0, y_1, x)}_{\text{max penalty}}\Bigg] \quad (30)
\end{aligned}
$$

We can maximize this unconstrained objective to find conditionally optimal dual variables, as stated below.

Proposition 4.1. *Suppose $\hat{\nu}_{0,x}^{\text{init}}, \hat{\nu}_{1,x}^{\text{init}} : \mathcal{Y} \to \mathbb{R}$ and $\hat{\lambda}_{x,1}, \ldots, \hat{\lambda}_{x,L} \geq 0$ maximize the objective $O(\nu_{0,x}, \nu_{1,x}, \{\lambda_{x,\ell}\}_{\ell=1}^{L})$ among all functions $\nu_{0,x}, \nu_{1,x} : \mathcal{Y} \to \mathbb{R}$ and constants $\lambda_{x,1}, \ldots, \lambda_{x,L} \geq 0$. Let c_x be the minimum constant such that $(\hat{\nu}_{0,x}^{\text{init}} - c_x, \hat{\nu}_{1,x}^{\text{init}} - c_x) \in \mathcal{V}_x$ are conditionally valid dual variables. Then $\hat{\nu}_{0,x}^{\text{init}} - c_x, \hat{\nu}_{1,x}^{\text{init}} - c_x$ solve the conditional dual problem Eq. (28).*

In other words, if we can find initial solutions $\hat{\nu}_{0,x}^{\text{init}}, \hat{\nu}_{1,x}^{\text{init}}, \hat{\lambda}_{x,1}, \ldots, \hat{\lambda}_{x,L} = \arg\max O(\nu_{0,x}, \nu_{1,x}, \{\lambda_{x,\ell}\}_{\ell=1}^{L})$, we can simply apply the grid-search from Section 4.1 to find an optimal solution to the conditional dual problem. In practice, we recommend optimizing a sample version of this objective. In particular, let $\{\tilde{Y}_b(0), \tilde{Y}_b(1)\}_{b=1}^{B}$ denote samples from $\hat{P}_{\text{prod},x}$ for some large B. Then equation (30) can be approximated by

$$
\begin{aligned}
\hat{O}(\nu_{0,x}, \nu_{1,x}, \{\lambda_{x,\ell}\}_{\ell=1}^{L}) = \frac{1}{B}\sum_{b=1}^{B}\Bigg\{&\nu_{0,x}(\tilde{Y}_b(0)) + \nu_{1,x}(\tilde{Y}_b(1)) \\
&- \max_{b \in [B]}\left[\nu_{0,x}(\tilde{Y}_b(0)) + \nu_{1,x}(\tilde{Y}_b(1)) - \sum_{\ell=1}^{L} \lambda_{x,\ell} w_{x,\ell}(\tilde{Y}_b(0), \tilde{Y}_b(1)) - f(\tilde{Y}_b(0), \tilde{Y}_b(1), x)\right]\Bigg\}.
\end{aligned}
$$

The sub-gradient with respect to $\alpha_{k,m,x}$ can be easily computed and hence it can optimized via gradient-based methods.

One shortcoming of the above approach is that the objective function is non-smooth. An alternative strategy is to use a smooth approximation of the exact objective equation (30):

$$
O_\epsilon(\nu_{0,x}, \nu_{1,x}, \{\lambda_{x,\ell}\}_{\ell=1}^{L}) \coloneqq \mathbb{E}_{\hat{P}_{\text{prod},x}}[\nu_{0,x}(Y(0)) + \nu_{1,x}(Y(1)) - R_{\epsilon,x}(Y(1), Y(0))]
$$

where the random variable $R_\epsilon(Y(1), Y(0))$ is the following smoothed penalty function:

$$
R_{\epsilon,x}(Y(1), Y(0)) = \epsilon \exp\left(\frac{\nu_{0,x}(Y(0)) + \nu_{1,x}(Y(1)) - \sum_{\ell=1}^{L} \lambda_{x,\ell} w_{x,\ell}(Y(0), Y(1)) - f(Y(0), Y(1), x)}{\epsilon}\right).
$$

This smooth penalty is typically known as an entropy regularizer in optimal transport theory (Villani et al., 2009; Peyré et al., 2019). Note for each ϵ, using the basis approximation $\nu_{k,x}(y) \approx \sum_{m=1}^{M} \alpha_{k,m,x}\phi_m(y)$, maximizing $O_\epsilon(\nu_{0,x}, \nu_{1,x}, \{\lambda_{x,\ell}\}_{\ell=1}^{L})$ is now a finite-dimensional unconstrained concave problem which we can solve using stochastic gradient descent. Thus, as a heuristic algorithm (which is commonly used in the optimal transport literature), we suggest using stochastic gradient descent to maximize the smoothed objective and sending $\epsilon \to 0$ along some schedule as we take more gradient steps. This algorithm is closely related to the Sinkhorn algorithm, and indeed, this optimization strategy is widely used in optimal transport literature (e.g. Sinkhorn, 1964; Villani et al., 2009; Cuturi, 2013; Altschuler et al., 2017; Peyré et al., 2019).

The main message is as follows: since the conditional problem Eq. (28) is only optimizing over two univariate functions, the literature contains many strategies to solve it approximately, including many additional methods beyond the two in this section. When combined with the general strategy outlined in Section 4.1, any of these methods can be used to compute the estimated dual variables $\hat{\nu}$. Crucially, as long as we effectively perform the two-dimensional grid search in Section 4.1, we will get valid bounds on θ_L no matter how poorly we solve Eq. (28).

5 Empirical applications

We now illustrate our method in applications to two randomized experiments and one observational study. Code and data are publicly available at `https://github.com/amspector100/dual_bounds_paper`.

5.1 Persuasion effects of political news

We first analyze data from Gerber et al. (2009), who in 2005 randomly assigned a set of individuals in Prince William County, Virginia, to receive a free subscription offer for the Washington Post.[5] Using administrative data, they also determined whether each subject voted in the November 2006 elections. Thus, for $n = 2400$ individuals, $W_i \in \{0, 1\}$ denotes whether individual i received a free subscription to the Washington Post, and $Y_i \in \{0, 1\}$ denotes whether individual i voted in the 2006 elections.

In this context, Jun and Lee (2023) (henceforth JL) studied the "persuasion effect" of the treatment, defined as the probability that the treatment causes an individual who would not otherwise have voted to vote:

$$\theta(P^\star) \coloneqq P^\star(Y(1) = 1 \mid Y(0) = 0) = \frac{P^\star(Y(1) = 1, Y(0) = 0)}{P^\star(Y(0) = 0)}. \tag{31}$$

This estimand is also known as the Probability of Sufficiency Pearl (1999). As noted by JL, without covariates, the sharp bounds on $\theta(P^\star)$ are rescaled Fréchet-Hoeffding bounds:

$$\theta_L^{\text{no-covariates}} \coloneqq \max\left(\frac{\mathbb{E}_{P^\star}[Y(1) - Y(0)]}{1 - \mathbb{E}_{P^\star}[Y(0)]}, 0\right) \leq \theta(P^\star) \leq \min\left(\frac{\mathbb{E}_{P^\star}[Y(1)]}{1 - \mathbb{E}_{P^\star}[Y(0)]}, 1\right) \coloneqq \theta_U^{\text{no-covariates}}.$$

However, Gerber et al. (2009) also collected a rich set of covariate information, including demographic information, political preferences, and previous voter turnout. Furthermore, $\theta(P^\star)$ takes the form of an unidentifiable expectation divided by an identifiable expectation (since $P^\star(Y(0) = 1)$ is identified—see Eq. 31). Thus, we can use our methodology to form covariate-assisted estimates of the numerator and apply the bivariate delta method to perform inference on $\theta(P^\star)$, as described in Appendix A.2.

To form the dual bounds, we estimate the conditional laws of $Y(1) \mid X$ and $Y(0) \mid X$ using three outcome models: a cross-validated logistic ridge regression, a random forest, and a k-nearest neighbors (KNN) classifier, where the covariates are the 43 baseline covariates from Gerber et al. (2009) plus interaction terms with the treatment. For each outcome model, we form dual bounds following the methodology from Sections 3.3-4 using 10-fold cross-fitting. We also compute non-robust plug-in bounds, which plug in the estimated conditional distributions and the empirical law of X into Eq. (**??**); unlike dual bounds, these bounds can be anti-conservatively biased. We also aggregate the results across all dual bounds using the multiplier bootstrap-like procedure detailed in Appendix A.2.

Table 1 shows the results, from which we report three main findings. First, the covariate-assisted dual bounds are more than twice as narrow as the covariate-free bounds. Second, the dual bounds appear to be more reliable than the covariate-assisted plug-in bounds. For example, the KNN and random forest outcome models produce plug-in lower bounds larger than 15%. This is implausible because the ATE point estimate is 0.029 and not significant; indeed, we do not even have power to reject the sharp null that $Y(1) = Y(0)$ with probability one. In contrast, all three outcome models provide dual bounds that are provably valid without assuming that the outcome model is accurate. Third and finally, the multiplier bootstrap method successfully selects the tightest lower and upper bounds while providing rigorous uncertainty quantification.

Remark 8. Our analysis is inspired by JL, but it differs from theirs in three ways. First, JL do not leverage covariates in their main empirical results. Second, JL assume that $Y(1) \geq Y(0)$ holds a.s. We chose to avoid this assumption, since prior work has shown that media exposure can sometimes depress turnout (e.g., Gentzkow, 2006), suggesting the treatment effect may be heterogeneous even if it is positive on average. Lastly, JL perform an instrumental variables (IV) analysis where the exposure is whether an individual *read* the Washington Post and the outcome is whether an individual voted for a Democrat. However, their exposure and outcome were only collected for $\approx 30\%$ of the sample who responded to a follow-up survey; thus, by performing an ITT analysis with voter turnout as the outcome, we avoid any missing data problems. It is possible to extend our methodology to IV analyses, but it requires new methodological ideas which we defer to a separate work (CITE TODO).

5.2 Estimating intensive margins

Carranza et al. (2022) conducted a randomized experiment in South Africa where treated individuals received assessment results that they could share with potential employers. They found that treated individuals had

[5] The original experiment had a third treatment condition, namely to receive a free subscription offer for the Washington Times. For simplicity, we follow Jun and Lee (2023) and only analyze subjects in the Washington Post or control treatment groups.

Outcome model	R^2	Dual LB	Dual UB	Plug-in LB	Plug-in UB
No covariates	0.0	0.056 (0.04) [0.0]	0.966 (0.039) [1.0]	0.057 (0.034) [0.0]	0.966 (0.031) [1.0]
Ridge	0.49	0.038 (0.027) [0.0]	0.365 (0.019) [0.403]	0.046	0.376
RF	0.365	0.003 (0.022) [0.0]	0.41 (0.021) [0.451]	0.158	0.348
KNN	0.358	0.0 (0.019) [0.0]	0.409 (0.021) [0.45]	0.192	0.366
Multiplier bootstrap*	-	0.056 [0.0]	0.365 [0.412]		

Table 1: This table shows the lower and upper bounds on $P^\star(Y(1) = 1 \mid Y(0) = 0)$ for the experimental data from Gerber et al. (2009), with standard errors shown in parentheses and confidence bounds shown in brackets. We do not know how to compute standard errors for the covariate-assisted plug-in bounds, so we do not list them. *Note that we do not use the exact multiplier bootstrap methodology from Section 2.3. Rather, we use the variant from Appendix A.2, which permits the use of cross-fitting.

higher employment rates and higher earnings, suggesting that the tests provided useful information about workers' skills. However, we might wonder: is the treatment effect driven by increases in employment (extensive margin), or does the treatment increase hours worked for individuals who would have been employed with or without the treatment (intensive margin)?

To estimate the intensive margin, Chen and Roth (2023) (henceforth CR) analyzed the following quantities:

$$\mathbb{E}[Y(1) - Y(0) \mid Y(1) > 0, Y(0) > 0] \text{ and } \mathbb{E}\left[\log(Y(1)) - \log(Y(0)) \mid Y(1) > 0, Y(0) > 0\right],$$

where above, the outcome Y measures the average hours worked per week post-treatment, and the logs in the latter estimand ensure that it is scale-invariant and can roughly be interpreted as a "percentage" effect. CR bounded these quantities using the methodology from Lee (2009), which assumes that $Y(1) > 0$ holds whenever $Y(0) > 0$, i.e., the treatment does not cause any individual to be unemployed. To defend this assumption, CR noted that individuals with poor test results likely did not share them with their employers, and we agree that this assumption seems plausible in this setting.

However, the dataset from Carranza et al. (2022) contains a rich set of pre-treatment covariates, including baseline earnings, demographic information, and educational history. Thus, we produce covariate-assisted variants of the bounds from CR. To fit the outcome model, we use the default settings in the `dualbounds` package, which employs a linear model with interactions:

$$Y_i = X_i^T \beta + W_i X_i^T \gamma + \epsilon_i. \tag{32}$$

To estimate β and γ, we use a cross-validated ridge regression. We estimate the law of $\epsilon_i \mid X_i, W_i$ as the empirical law of the estimated residuals $\{Y_i - X_i^T\hat{\beta} - W_i X_i^T\hat{\gamma} : i \in [n], W_i = w\}$.[6] We then convert this outcome model into a cross-fit dual bound using the methodology from Sections 2 and 4.

Table 2 shows the results: for both the logged and non-logged outcome, the covariate-assisted bounds are only $\approx 60\%$ as wide as the covariate-free bounds. Although the bounds are still quite wide, this analysis nonetheless shows that covariate adjustment can substantially sharpen partial identification bounds without requiring additional assumptions.

5.3 401k eligibility

We now study how 401(k) eligibility impacts wealth. An extensive literature argues that 401(k) eligibility is essentially exogenous conditional on covariates (e.g., Poterba et al., 1995; Poterba and Venti, 1998; Poterba et al., 2000; Chernozhukov and Hansen, 2004), since workers likely choose employers based on job characteristics besides 401(k) eligibility, e.g., income. We adopt this assumption; thus, the outcome $Y \in \mathbb{R}$ measures total household wealth and the treatment $W \in \{0, 1\}$ indicates 401(k) eligibility.

[6] This estimate severely restricts the heteroskedasticity pattern, since it asserts that the residuals are independent of the covariates given the treatment, i.e., $\epsilon_i \perp\!\!\!\perp X_i \mid W_i$. That said, we emphasize that the final dual bounds are valid even if the model for the law of $\epsilon_i \mid X_i, W_i$ is completely inaccurate.

	Log-hours		Hours	
Outcome model	Lower bound	Upper bound	Lower bound	Upper bound
No covariates (plug-in)	-0.193	0.281	-6.64	2.69
	(0.062)	(0.111)	(1.36)	(2.06)
Ridge (dual)	-0.115	0.185	-4.74	1.18
	(0.060)	(0.130)	(1.46)	(2.00)

Table 2: This table shows the lower and upper bounds on $\mathbb{E}[\log(Y(1)) - \log(Y(0)) \mid Y(1) > 0, Y(0) > 0]$ and $\mathbb{E}[Y(1)) - Y(0) \mid Y(1) > 0, Y(0) > 0]$ for the dataset from Carranza et al. (2022). Standard errors are shown in parentheses and clustered at the assessment date level (following Chen and Roth (2023)).

We obtain data from Chernozhukov et al. (2018a), who estimated average treatment effects (ATEs) using a sample of households from the 4th wave of the 1990 Survey of Income and Program Participation. Yet the literature emphasizes that treatment effects may be highly heterogeneous. For instance, 401(k) eligibility may reduce wealth for households who otherwise would participate in a different retirement plan or whose 401(k) contributions adversely reduce their liquidity. To study whether negative effects contribute to the overall ATE, we now bound the *positive treatment effect*:

$$\theta(P^\star) = \mathbb{E}_{P^\star}\left[\max(Y(1) - Y(0), 0)\right]. \tag{33}$$

Our analysis uses the same covariates as Chernozhukov et al. (2018a), including income, demographics, and financial indicators such as homeownership status. Following Chernozhukov et al. (2018a), we use the raw covariates except when fitting regularized GLMs, where we include polynomial transformations and pairwise interactions We estimate cross-fit propensity scores using a cross-validated logistic elastic net.

The outcome model takes the form $Y_i = \mathbb{E}[Y_i \mid X_i, W_i] + \epsilon_i$. To estimate $\mathbb{E}[Y_i \mid X_i, W_i]$, we use (i) a cross-validated elastic net, (ii) a KNN regressor, and (iii) an `sklearn` histogram gradient boosting (HGBoost) regressor (Pedregosa et al., 2011; Ke et al., 2017) with the constraint that $\mathbb{E}[Y_i \mid X_i, W_i]$ is increasing in W_i.[7] We estimate the law of $\epsilon_i \mid X_i, W_i$ as in Section 5.2. For each model, we report cross-fit AIPW dual bounds (as described in Sections 3.4-4) as well as non-robust "plug-in" bounds, which plug $\hat{P}_{Y|X,W}$ and the empirical law of the covariates into Eq. (**??**).

Method	R^2	ATE	Dual LB	Dual UB	Plug-in LB	Plug-in UB
HGBoost	0.4255	6381	5564	47286	5834	51075
		(1882)	(1201)	(1258)		
KNN	0.3776	6792	4476	47424	10135	45800
		(1966)	(1640)	(1333)		
Elastic net	0.1672	10637	6938	60940	9078	55005
		(2284)	(2091)	(1579)		

Table 3: This table shows estimated average treatment effects as well as lower and upper bounds on $\mathbb{E}[\max(Y(1) - Y(0), 0)]$ for the 401(k) eligibility dataset. Standard errors are shown in parentheses. We do not know how to compute standard errors for the covariate-assisted plug-in bounds, so we do not list them.

Table 3 shows the out-of-sample R^2, cross-fit AIPW ATE estimates, dual bounds, and plug-in bounds for each outcome model. We report two main conclusions.

1. Incorporating covariates improves robustness. It is known that in observational studies, accurate outcome models can reduce the bias of ATE estimates. E.g., in Table 3, more accurate outcome models yield smaller ATE estimates, ranging from $\approx$ \$11K to $\approx$ \$6K. Table 3 suggests that the same logic applies to partially identified estimands (see Theorem 3.5), since the dual lower bounds decrease with the ATE estimates.

2. Plug-in bounds can be anticonservative. The KNN plug-in lower bound is $\approx$ \$10K. This value seems implausible, since it is twice as large as the corresponding dual bound and 50% larger than the ATE estimate from the best-performing model. Indeed, covariate-assisted plug-in bounds rely entirely on the

[7] Note that while the "HGBoost monotone" model asserts that the conditional average treatment effect $\mathbb{E}[Y(1) - Y(0) \mid X]$ is nonnegative, it nonetheless allows the positive treatment effect $\theta(P^\star) = \mathbb{E}[\max(Y(1) - Y(0), 0)]$ to differ from the ATE, for example, because there may be unobserved covariates U such that $\mathbb{E}[Y(1) - Y(0) \mid X, U]$ may be negative.

accuracy of the outcome model, whereas dual bounds are doubly robust as per Theorem 3.5. That said, it is reassuring that the best-performing model (HGBoost) yields similar plug-in and dual bounds.

Remark 9. Although the ATE lower bounds $\theta(P^\star)$, the dual lower bounds are smaller (0–2 standard errors) than the ATE estimates. This is a consequence of fitting an imperfect outcome model, leading to conservative bounds. That said, employing minimum norm optimal dual variables appears to reduce standard errors. E.g., for the HGBoost model, the lower confidence bound for $\theta(P^\star)$ is tighter than the ATE lower confidence bound. Nonetheless, this phenomenon is undesirable—we leave the problem of correcting it to future work.

5.4 A Monte-Carlo simulation

In this section, we run simulations to demonstrate the power, validity, and computational efficiency of dual bounds. Throughout, we consider randomized experiments where the propensity scores $\pi(x) = \frac{1}{2}$ are known.

We perform simulations where we estimate lower Lee bounds (Example 3). We sample covariates $X_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_p)$ for $p = 20$ covariates and draw $Y_i \mid X_i$ from a homoskedastic Gaussian linear model:

$$Y_i(1) \mid X_i \sim \mathcal{N}(X_i^T\beta + \tau, \sigma^2) \text{ and } Y_i(0) \mid X_i \sim \mathcal{N}(X_i^T\beta, \sigma^2) \tag{34}$$

for variance $\sigma^2 = 1$, coefficients $\beta \in \mathbb{R}^p$ chosen such that $\text{Var}(Y_i(1)) = \text{Var}(Y_i(0)) = 10$, and average treatment effect $\tau = 2$. We sample the selection events $S_i \mid X_i$ from a logistic regression model:

$$\mathbb{P}(S_i(0) = 1 \mid X_i) = \text{logit}^{-1}\left(X_i^T\beta_S + \tau_{S,0}\right) \text{ and } \mathbb{P}(S_i(1) = 1 \mid X_i) = \text{logit}^{-1}\left(X_i^T\beta_S + \tau_{S,1}\right) \text{ for } \beta_S \in \mathbb{R}^p \tag{35}$$

with $\|\beta_S\|_2 = 1$ and $\tau_{S,0} = 0, \tau_{S,1} = 1$. Following general practice in the literature, our simulations enforce the monotonicity condition $S(1) \geq S(0)$ a.s., and we assume that the practitioner knows this a-priori. We compare four methods for estimating the sharp bound θ_L in this problem:

1. The "naive plug-in" method first estimates $\hat{\beta}, \hat{\tau}, \hat{\beta}_S, \hat{\tau}_S$ using cross-validated ridge and logistic ridge regressions, and we estimate $\hat{\sigma}$ as the sample standard deviation of the estimated residuals $\{Y_i - X_i^T\hat{\beta}\}_{i\in[n]}$. Then, we approximate the law of $Y_i(k) \mid X_i$ and $S_i(k) \mid X_i$ by plugging in the estimated values of $\hat{\beta}, \hat{\tau}, \hat{\beta}_S, \hat{\tau}_S$ and $\hat{\sigma}$ to Equations (34) and (35). At this point, we can plug the estimated laws of $Y_i(k) \mid X_i, S_i(k) \mid X_i$ into the formula for θ_L (see Eq. (5)), yielding an estimate $\hat{\theta}_L^{\text{plugin}}$. In general, it is not clear how to compute standard errors for $\hat{\theta}_L^{\text{plugin}}$; to be as generous as possible, we compute oracle lower confidence bounds using the true variance

$$\hat{\theta}_{\text{LCB}}^{\text{plugin}} = \hat{\theta}_L^{\text{plugin}} - \Phi^{-1}(1-\alpha)\sqrt{\text{Var}\left(\hat{\theta}_L^{\text{plugin}}\right)}. \tag{36}$$

   We compute the true value of $\text{Var}\left(\hat{\theta}_L^{\text{plugin}}\right)$ numerically by sampling many datasets from the true data-generating process.

2. The "dual crossfit" approach uses *exactly* the same approach to estimate the conditional laws $Y_i(k) \mid X_i$ and $S_i(k) \mid X_i$ (with the exception that it employs cross-fitting). However, after computing the estimates of these laws on $K = 5$ folds of the data, we apply the cross-fit dual bounds methodology from Section 3.3. Since Y is continuous, to compute the estimated dual variables $\hat{\nu}$, we use the discretization approach outlined in Section 4.2 with $n_{\text{vals}} = 50$ discretizations.[8] Computing dual bounds using this method takes less than 5 seconds with $n = 1000$ observations in our simulations.

3. The "dual oracle" approach applies dual bounds but uses the true optimal dual variables $\nu^\star$ based on the true data generating process of the simulation. Note the dual oracle estimator $\hat{\theta}_L^{\star\star}$ and LCB $\hat{\theta}_{\text{LCB}}^{\star\star}$ are defined explicitly in Eq. (23).

4. The "no covariates" method is identical to the naive plug-in approach except that it does not observe the covariates and only estimates the marginal laws of $Y_i(1), Y_i(0), S_i(1), S_i(0)$.

We also consider the performance of each method in two misspecified settings where $Y_i \mid X_i$ is actually heteroskedastic:

$$Y_i(1) \mid X_i \sim \mathcal{N}(X_i^T\beta + \tau, \sigma_1^2(X_i)) \text{ and } Y_i(0) \mid X_i \sim \mathcal{N}(X_i^T\beta, \sigma_0^2(X_i)) \tag{37}$$

[8] We remind the reader that the final dual lower confidence bound will be valid no matter how small n_{vals} is, although increasing n_{vals} may yield higher power.

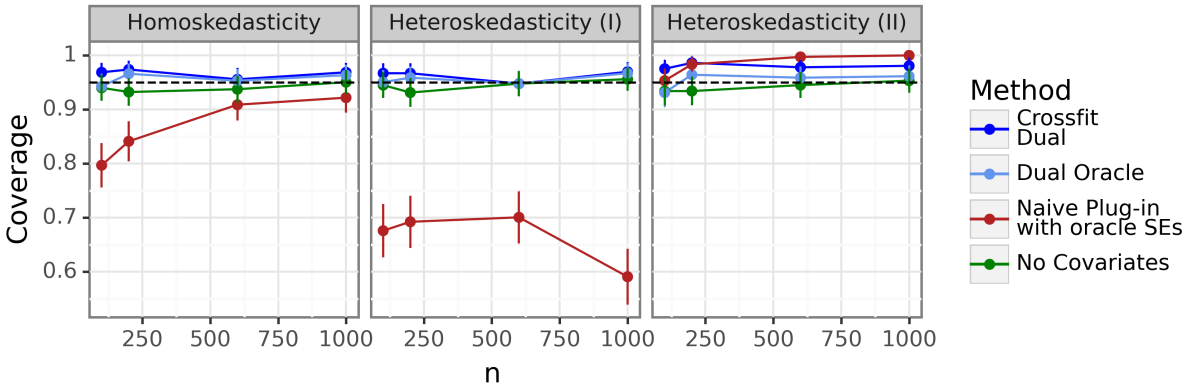

Figure 2: This figure shows the coverage of the lower Lee confidence bounds from Figure 1. The nominal level is 95%, shown by the dotted black line.

for the functions $\sigma_1^2(X) = \sigma_1^2\|X\|_2^2, \sigma_0^2(X) = \sigma_0^2\|X\|_2^2$ for constants $\sigma_1, \sigma_0 \geq 0$. In this case, for both the naive plug-in and dual crossfit methods, the estimated outcome model is misspecified, since it incorrectly assumes homoskedasticity. In the first setting (labelled as "Heteroskedasticity (I)"), we set $\sigma_1/\sigma_0 = 3$; in the other setting ("Heteroskedasticity (II)"), we set $\sigma_0/\sigma_1 = 0.3$.

Figures 1,2, and 3 show the results with $n \in \{100, 200, 600, 1000\}$. Figure 1 shows the average value of the estimate $\hat{\theta}_L$ and the lower confidence bound $\hat{\theta}_{\text{LCB}}$; it shows that the naive-plug in estimator is biased when n is small (due to the effect of regularization) and when the model is misspecified. In contrast, the cross-fit dual bounds are (i) guaranteed to be conservatively biased at worst and (ii) less sensitive to errors in estimating the outcome model, yielding valid and reasonably sharp inference in all three settings. Figure 2 confirms that dual bounds provide $\geq 95\%$ coverage in all settings, whereas the naive plug-in method can be quite conservative or anticonservative, depending on the form of heteroskedasticity. Lastly, to further analyze the tightness of our method, Figure 3 displays the distribution of the cross-fit and oracle dual bounds. As predicted by our theory (e.g. Prop. 3.1), for large n, the cross-fit dual bounds are nearly indistinguishable from the oracle bounds. Surprisingly, this result also roughly holds even for $n = 100$, which is remarkable given that there are only twice as many observations as nuisance parameters in this case. Overall, these results suggest that at least in this setting, cross-fit dual bounds are near optimal even in moderate dimensions.

6 Discussion

This paper introduced a method to compute confidence intervals for a class of partially identified parameters in econometrics and causal inference. These confidence intervals can leverage modern techniques from machine learning to narrow the confidence intervals by learning the relationship between the covariates X and the outcome Y; however, in randomized experiments, the confidence intervals are still robust to arbitrary misspecification of the outcome model. Furthermore, when the model is well-specified, these confidence intervals are asymptotically equivalent to an oracle confidence interval formed using perfect knowledge of the joint laws $(Y(0), X)$ and $(Y(1), X)$.

However, our analysis leaves open many questions. For example, what is the best way to estimate the optimal dual variables $\nu^\star$? We give a few suggestions in this paper with appealing properties, but any choice of $\hat{\nu}$ will lead to valid inference, and other choices may lead to higher power or faster computation. Additionally, a few of our theoretical results require Y to be discrete, and it would be interesting to investigate if the same results hold when Y is continuous.

Perhaps the most pressing question is whether the techniques developed in this paper can be applied more generally. In particular, we use an extremely simple duality argument to guarantee that our method is robust to arbitrary misspecification. Does this same argument apply to settings beyond causal inference? In the next two sections, we begin to address this question.

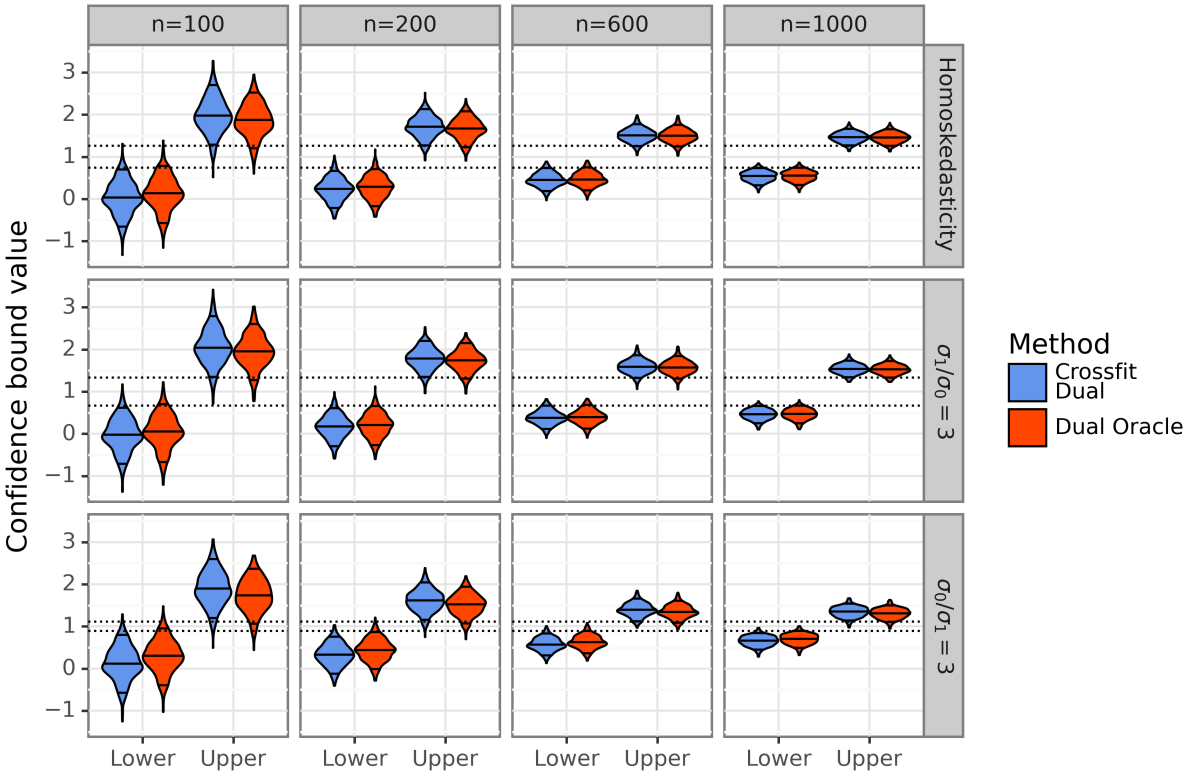

Figure 3: This figure shows the distribution of the lower and upper confidence bounds from the cross-fit dual and dual oracle methods. Note that the dotted lines denote θ_L and θ_U, the true bounds, and within each density plot, the three lines denote the $0.05, 0.5$, and 0.95 quantiles.

6.1 Extensions beyond causal effects

In this section, we discuss whether our core method can be extended to settings that cannot be reduced to estimating an expectation of the form $\mathbb{E}[f(Y(0), Y(1), X)]$. Indeed, the ideas in Section 2.2 apply to many classes of estimands which are defined as the solution to optimization problems. There are countless problems of this form in economics and statistics (Chernozhukov et al., 2007); we describe two classes of problems below.

Example 7 (Inference for linear programs)**.** Suppose that θ is the optimal objective value of a linear program of the form

$$\theta := \min_{z \in \mathbb{R}^d} c^T z \text{ s.t. } Az \leq b(P),$$

where $A \in \mathbb{R}^{d \times m}$ is a known matrix and $b(P) \in \mathbb{R}^d$ is a vector of moments or conditional moments of a probability distribution P. Many estimands in economics can be written this way (e.g. Gafarov, 2019; Fang et al., 2023), including those arising in models of demand (Tebaldi et al., 2023; Nevo et al., 2016) and income mobility (Chetty et al., 2017). If we observe i.i.d. samples from P, the exact same method from Section 2.2 can be applied to obtain a $1 - \alpha$ lower confidence bound on θ.

Example 8 (Variance of the CATE)**.** In Example 2, we noted that $\mathrm{Var}(Y(1) - Y(0))$ is a natural measure of treatment effect heterogeneity. Another interesting estimand is the variance of the conditional average treatment effect (CATE) $\tau(X) := \mathbb{E}[Y(1) - Y(0) \mid X]$. Using Fenchel conjugacy (or Cauchy-Schwartz), we can derive a dual representation:

$$\begin{aligned}\mathrm{Var}(\tau(X)) &= \max_{h:\mathcal{X}\to\mathbb{R}} 2\,\mathrm{Cov}(h(X), \tau(X)) - \mathrm{Var}(h(X)) \\ &= \max_{h:\mathcal{X}\to\mathbb{R}} 2\,\mathrm{Cov}(h(X), Y(1) - Y(0)) - \mathrm{Var}(h(X)).\end{aligned}$$

Crucially, the bound $B(h) := 2\,\mathrm{Cov}(h(X), Y(1) - Y(0)) - \mathrm{Var}(h(X))$ is easy to estimate (in randomized experiments) for any fixed $h : \mathcal{X} \to \mathbb{R}$; thus, we can obtain a robust lower confidence bound on $\mathrm{Var}(\tau(X))$ by selecting a function $\hat{h} \approx \arg\max_h B(h)$ on the first split of data and estimating $B(\hat{h})$ on the second split.

This idea is connected to Zhang and Janson (2020); Wang et al. (2023), who also select a lower bound for nonparametric variance estimation using a different variational representation. However, the insight of our approach is that it uses duality to select an easy-to-estimate lower bound, yielding misspecification-robust inference.

6.2 A counterexample: robustness when the propensity scores are unknown

However, linear dual bounds as defined in Section 2.2 are not appropriate for every problem. For example, one might hope that a simple modification of Definition 1 can produce "always valid" bounds on the average treatment effect when the propensity scores $\pi(X_i)$ are not known. Unfortunately, this is not the case, as shown by the following counterexample.

Counterexample 1 (Average treatment effect with unknown propensity scores)**.** Suppose the vectors $(X_i, W_i, Y_i(0), Y_i(1))$ are sampled i.i.d. from some population distribution $P^\star \in \mathcal{P}$, where $\mathcal{P}$ is the set of distributions on $\mathcal{X} \times \{0,1\} \times \mathcal{Y}^2$ satisfying unconfoundedness, i.e., $\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp W_i \mid X_i$, and strict overlap, i.e., $\inf_{x\in\mathcal{X}} \pi_P(x) > 0$ and $\sup_{x\in\mathcal{X}} \pi_P(x) < 1$ for all $P \in \mathcal{P}$, where $\pi_P(x) \coloneqq \mathbb{E}_P[W \mid X = x]$.

Given i.i.d. observations (X_i, W_i, Y_i), we seek to form a lower bound on the average treatment effect $\theta(P^\star) \coloneqq \mathbb{E}_{P^\star}[Y_i(1) - Y_i(0)]$ which is valid even under arbitrary misspecification of $\pi_P(X_i)$ and the outcome model. Although θ is identifiable, it can still be written as the solution to the optimization problem

$$\theta(P^\star) = \min_{P\in\mathcal{P}} \mathbb{E}_P[Y(1) - Y(0)] \text{ s.t. } P_{X,W,Y} = P^\star_{X,W,Y} \tag{38}$$

where $P_{X,W,Y}$ is the law of (X, W, Y) under P and $P^\star_{X,W,Y}$ is the true law of (X, W, Y). Note that the optimization variable is P, which is a joint law over $(X, W, Y(0), Y(1))$, and $P_{X,W,Y}$ is a functional of P. For any $h : \mathcal{X} \times \{0,1\} \times \mathcal{Y} \to \mathbb{R}$, the Lagrange dual to this problem is

$$g(h) \coloneqq \mathbb{E}_{P^\star}\left[h(X, W, Y)\right] + \kappa(h),$$

where $\kappa(h) \coloneqq \inf_{P\in\mathcal{P}} \mathbb{E}_P[Y(1) - Y(0) - h(X, W, Y)]$ is a known constant depending on h. For any h, $g(h)$ is a valid lower bound on $\theta(P^\star)$ by weak duality, but unfortunately, strong duality does not hold. In particular, for any h, we have that

$$h(X_i, W_i, Y_i) + \kappa(h) \le \begin{cases} Y_i - \max(\mathcal{Y}) & W_i = 1 \\ \min(\mathcal{Y}) - Y_i & W_i = 0. \end{cases}$$

See Appendix E for a proof.

Counterexample 1 tells us that any dual bound (in the sense of Def. 1) on the ATE which is valid under arbitrary misspecification must also be trivial, since it must impute $Y_i(1)$ to have the minimum possible value whenever it is not observed, and it must impute $Y_i(0)$ to have the maximum possible value when it is not observed. This result is very negative, because it holds even in the case where $\mathcal{X}$ contains only one element and thus the ATE can be consistently estimated by the standard difference-in-means estimator:

$$\hat{\theta}_{\text{diff}-\text{in}-\text{means}} \coloneqq \frac{\sum_i W_i Y_i}{\sum_i W_i} - \frac{\sum_i (1 - W_i) Y_i}{\sum_i 1 - W_i}.$$

In other words, when the propensity scores are unknown, a naive application of dual bounds will require the estimator to be written as a sample mean, but this class of estimators is too restrictive. Another interpretation is that strong duality does not hold when the propensity scores are unknown. On the other hand, Aronow et al. (2021) prove that no uniformly consistent estimator of ATE exists under strong ignorability and strict overlap without further assumptions if one of the covariates is continuous. The counterexample confirms this result for Dual Bounds.

6.3 Improving the computational strategy

To compute our recommended estimator $\hat{\nu}$ of the optimal dual variables, one must first model the conditional laws $P^\star_{Y(0)|X}, P^\star_{Y(1)|X}$. This procedure may not yield tight confidence intervals when conditional distributions are hard to model. For example, when X includes structured data such as images and texts, simple machine learning algorithms are unable to provide accurate distribution estimates. In Appendix B, we present Deep

Dual Bounds, an alternative approach that directly learn the optimal dual variables $\hat{\nu}_0, \hat{\nu}_1$ by solving the optimal transport problem in the product space $\mathcal{Y} \times \mathcal{X}$:

$$\theta_L = \sup_{\nu_0, \nu_1 \in \mathcal{V}} \mathbb{E}_{P^\star}[\nu_0(Y(0), X) + \nu_1(Y(1), X)]. \tag{39}$$

The core idea of this approach is to incorporate the conditional distribution estimation step into the optimization objective, and therefore we can directly solve the optimal dual variables as a function of both covariates X and outcome Y without decomposition. This direct approach involves only a single optimization problem, which enables us to use deep neural networks to model the high dimension dual variables $\nu_0(Y(0), X), \nu_0(Y(0), X)$, and optimize with gradient based method in an end-to-end fashion. The Deep Dual Bounds can deal with unstructured covariates like texts, where direct modeling of conditional distributions is often challenging. The details and experimental results of the algorithm are discussed in Appendix B and Table 4.

References

Abadie, A. and Imbens, G. W. (2006). Large sample properties of matching estimators for average treatment effects. *econometrica*, 74(1):235–267.

Adams, D. R. and Hedberg, L. I. (1999). *Function spaces and potential theory*, volume 314. Springer Science & Business Media.

Altschuler, J., Niles-Weed, J., and Rigollet, P. (2017). Near-linear time approximation algorithms for optimal transport via sinkhorn iteration. *Advances in neural information processing systems*, 30.

Andrews, D. W. K. and Shi, X. (2013). Inference based on conditional moment inequalities. *Econometrica*, 81(2):609–666.

Andrews, I., Kitagawa, T., and McCloskey, A. (2019). Inference on winners. Working Paper 25456, National Bureau of Economic Research.

Andrews, I., Roth, J., and Pakes, A. (2023). Inference for Linear Conditional Moment Inequalities. *The Review of Economic Studies*. rdad004.

Arkhangelsky, D., Imbens, G. W., Lei, L., and Luo, X. (2021). Double-robust two-way-fixed-effects regression for panel data. *arXiv preprint arXiv:2107.13737*.

Aronow, P., Robins, J. M., Saarinen, T., Sävje, F., and Sekhon, J. (2021). Nonparametric identification is not enough, but randomized controlled trials are. *arXiv preprint arXiv:2108.11342*.

Aronow, P. M., Green, D. P., and Lee, D. K. K. (2014). Sharp bounds on the variance in randomized experiments. *The Annals of Statistics*, 42(3):850–871.

Balke, A. and Pearl, J. (1997). Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association*, 92(439):1171–1176.

Bogachev, V. I. (2022). Kantorovich problems with a parameter and density constraints. *Siberian Mathematical Journal*, 63(1):34–47.

Bogachev, V. I. and Malofeev, I. I. (2020). Kantorovich problems and conditional measures depending on a parameter. *Journal of Mathematical Analysis and Applications*, 486(1):123883.

Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press, USA.

Carranza, E., Garlick, R., Orkin, K., and Rankin, N. (2022). Job search and hiring with limited information about workseekers' skills. *American Economic Review*, 112(11):3547–83.

Chandrasekhar, A., Chernozhukov, V., Molinari, F., and Schrimpf, P. (2012). Inference for best linear approximations to set identified functions.

Chen, J. and Roth, J. (2023). Logs with Zeros? Some Problems and Solutions*. *The Quarterly Journal of Economics*, 139(2):891–936.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018a). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68.

Chernozhukov, V., Chetverikov, D., and Kato, K. (2013a). Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *The Annals of Statistics*, 41(6):2786 – 2819.

Chernozhukov, V., Chetverikov, D., and Kato, K. (2018b). Inference on Causal and Structural Parameters using Many Moment Inequalities. *The Review of Economic Studies*, 86(5):1867–1900.

Chernozhukov, V., Fernández-Val, I., and Galichon, A. (2010). Quantile and probability curves without crossing. *Econometrica*, 78(3):1093–1125.

Chernozhukov, V., Fernández-Val, I., and Melly, B. (2013b). Inference on counterfactual distributions. *Econometrica*, 81(6):2205–2268.

Chernozhukov, V. and Hansen, C. (2004). The effects of 401(k) participation on the wealth distribution: An instrumental quantile regression analysis. *The Review of Economics and Statistics*, 86(3):735–751.

Chernozhukov, V., Hong, H., and Tamer, E. (2007). Estimation and confidence regions for parameter sets in econometric models. *Econometrica*, 75(5):1243–1284.

Chernozhukov, V., Lee, S., and Rosen, A. M. (2013c). Intersection bounds: Estimation and inference. *Econometrica*, 81(2):667–737.

Chernozhukov, V., Newey, W., Singh, R., and Syrgkanis, V. (2020). Adversarial estimation of riesz representers. *arXiv preprint arXiv:2101.00009*.

Chetty, R., Grusky, D., Hell, M., Hendren, N., Manduca, R., and Narang, J. (2017). The fading american dream: Trends in absolute income mobility since 1940. *Science*, 356(6336):398–406.

Chiburis, R. C. (2010). Semiparametric bounds on treatment effects. *Journal of Econometrics*, 159(2):267–275.

Cuturi, M. (2013). Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26.

Ding, P. and Lu, J. (2017). Principal stratification analysis using principal scores. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 79(3):757–777.

Dorn, J. and Guo, K. (2022). Sharp sensitivity analysis for inverse propensity weighting via quantile balancing. *Journal of the American Statistical Association*, 0(0):1–13.

Dorn, J., Guo, K., and Kallus, N. (2022). Doubly-valid/doubly-sharp sensitivity analysis for causal inference with unmeasured confounding.

Fan, Y., Guerre, E., and Zhu, D. (2017). Partial identification of functionals of the joint distribution of "potential outcomes". *Journal of econometrics*, 197(1):42–59.

Fan, Y. and Park, S. S. (2010). Sharp bounds on the distribution of treatment effects and their statistical inference. *Econometric Theory*, 26(3):931–951.

Fan, Y. and Park, S. S. (2012). Confidence intervals for the quantile of treatment effects in randomized experiments. *Journal of Econometrics*, 167(2):330–344.

Fang, Z., Santos, A., Shaikh, A. M., and Torgovitsky, A. (2023). Inference for large-scale linear systems with known coefficients. *Econometrica*, 91(1):299–327.

Firpo, S. and Ridder, G. (2008). Bounds on functionals of the distribution of treatment effects.

Firpo, S. and Ridder, G. (2019). Partial identification of the treatment effect distribution and its functionals. *Journal of Econometrics*, 213(1):210–234. Annals: In Honor of Roger Koenker.

Frangakis, C. E. and Rubin, D. B. (2002). Principal stratification in causal inference. *Biometrics*, 58(1):21–29.

Fréchet, M. (1951). Sur les tableaux de correlation dont les marges sont donnees. *Annals University Lyon: Series A*, pages 53–77.

Friedman, J. H. (2020). Contrast trees and distribution boosting. *Proceedings of the National Academy of Sciences*, 117(35):21175–21184.

Gafarov, B. (2019). Inference in high-dimensional set-identified affine models. *arXiv preprint arXiv:1904.00111*.

Gentzkow, M. (2006). Television and Voter Turnout*. *The Quarterly Journal of Economics*, 121(3):931–972.

Gerber, A. S., Karlan, D., and Bergan, D. (2009). Does the media matter? a field experiment measuring the effect of newspapers on voting behavior and political opinions. *American Economic Journal: Applied Economics*, 1(2):35–52.

Heckman, J. J., Smith, J., and Clements, N. (1997). Making the most out of programme evaluations and social experiments: Accounting for heterogeneity in programme impacts. *The Review of Economic Studies*, 64(4):487–535.

Hoeffding, W. (1940). Masstabinvariante korrelationstheorie. *Schriften des Mathematischen Instituts und des Institutes Fur Angewandte Mathematik der Universitiat Berlin*, pages 179–223.

Hoffman, A. J. (1952). On approximate solutions of systems of linear inequalities. *Journal of Research of the National Bureau of Standards*, 49(4).

Hsieh, Y.-W., Shi, X., and Shum, M. (2022). Inference on estimators defined by mathematical programming. *Journal of Econometrics*, 226(2):248–268.

Imbens, G. and Manski, C. (2004). Confidence intervals for partially identified parameters. *Econometrica*, 72(6):1845–1857.

Ipsen, I. C. F. and Nadler, B. (2009). Refined perturbation bounds for eigenvalues of hermitian and non-hermitian matrices. *SIAM Journal on Matrix Analysis and Applications*, 31(1):40–53.

Jun, S. J. and Lee, S. (2023). Identifying the effect of persuasion. *Journal of Political Economy*, 131(8):2032–2058.

Kaji, T. and Cao, J. (2023). Assessing heterogeneity of treatment effects.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Kneib, T., Silbersdorff, A., and Säfken, B. (2023). Rage against the mean – a review of distributional regression approaches. *Econometrics and Statistics*, 26:99–123.

Koenker, R. and Bassett Jr, G. (1978). Regression quantiles. *Econometrica: journal of the Econometric Society*, pages 33–50.

Lee, D. S. (2009). Training, wages, and sample selection: Estimating sharp bounds on treatment effects. *The Review of Economic Studies*, 76(3):1071–1102.

Lee, S. (2023). Partial identification and inference for conditional distributions of treatment effects.

Levis, A. W., Bonvini, M., Zeng, Z., Keele, L., and Kennedy, E. H. (2023). Covariate-assisted bounds on causal effects with instrumental variables. *arXiv preprint arXiv:2301.12106*.

Makkuva, A., Taghvaei, A., Oh, S., and Lee, J. (2020). Optimal transport mapping via input convex neural networks. In *International Conference on Machine Learning*, pages 6672–6681. PMLR.

Manski, C. F. (1990). Nonparametric bounds on treatment effects. *The American Economic Review*, 80(2):319–323.

Manski, C. F. (1997). Monotone treatment response. *Econometrica: Journal of the Econometric Society*, pages 1311–1334.

Manski, C. F. (2003). *Partial identification of probability distributions*, volume 5. Springer.

Manski, C. F. and Tamer, E. (2002). Inference on regressions with interval data on a regressor or outcome. *Econometrica*, 70(2):519–546.

Matoušek, J. and Gärtner, B. (2007). *Understanding and using linear programming*, volume 1. Springer.

Molinari, F. (2008). Partial identification of probability distributions with misclassified data. *Journal of Econometrics*, 144(1):81–117.

Molinari, F. (2020a). Microeconometrics with partial identification. *Handbook of econometrics*, 7:355–486.

Molinari, F. (2020b). Microeconometrics with partial identification.

Nevo, A., Turner, J. L., and Williams, J. W. (2016). Usage-based pricing and demand for residential broadband. *Econometrica*, 84(2):411–443.

Ober-Reynolds, D. (2023). Estimating functionals of the joint distribution of potential outcomes with optimal transport.

Page, L. C., Feller, A., Grindal, T., Miratrix, L., and Somers, M.-A. (2015). Principal stratification: A tool for understanding variation in program effects across endogenous subgroups. *American Journal of Evaluation*, 36(4):514–531.

Pearl, J. (1999). Probabilities of causation: Three counterfactual interpretations and their identification. *Synthese*, pages 93–149.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Peyré, G., Cuturi, M., et al. (2019). Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607.

Poterba, J. M. and Venti, S. F. (1998). Personal retirement saving programs and asset accumulation: Reconciling the evidence. In David A. Wise, e., editor, *Frontiers in the Economics of Aging*. National Bureau of Economic Research.

Poterba, J. M., Venti, S. F., and Wise, D. A. (1995). Do 401(k) contributions crowd out other personal saving? *Journal of Public Economics*, 58(1):1–32.

Poterba, J. M., Venti, S. F., and Wise, D. A. (2000). Saver behavior and 401(k) retirement wealth. *American Economic Review*, 90(2):297–302.

Ramdas, A. and Peña, J. (2016). Towards a deeper geometric, analytic and algorithmic understanding of margins. *Optimization Methods and Software*, 31(2):377–391.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866.

Robinson, S. M. (1973). Bounds for error in the solution set of a perturbed linear program. *Linear Algebra and its Applications*, 6:69–81.

Russell, T. M. (2021). Sharp bounds on functionals of the joint distribution in the analysis of treatment effects. *Journal of Business & Economic Statistics*, 39(2):532–546.

Semenova, V. (2021). Generalized lee bounds. *arXiv preprint arXiv:2008.12720*.

Semenova, V. (2023). Adaptive estimation of intersection bounds: a classification approach.

Semenova, V. and Chernozhukov, V. (2020). Debiased machine learning of conditional average treatment effects and other causal functions. *The Econometrics Journal*, 24(2):264–289.

Sinkhorn, R. (1964). A relationship between arbitrary positive matrices and doubly stochastic matrices. *The annals of mathematical statistics*, 35(2):876–879.

Stoye, J. (2009). More on confidence intervals for partially identified parameters. *Econometrica*, 77(4):1299–1315.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 25(1):1.

Sur, P. and Candès, E. J. (2019). A modern maximum-likelihood theory for high-dimensional logistic regression. 116(29):14516–14525.

Tamer, E. (2010). Partial identification in econometrics. *Annu. Rev. Econ.*, 2(1):167–195.

Tebaldi, P., Torgovitsky, A., and Yang, H. (2023). Nonparametric estimates of demand in the california health insurance exchange. *Econometrica*, 91(1):107–146.

Tetenov, A. (2012). Identification of positive treatment effects in randomized experiments with non-compliance.

Villani, C. et al. (2009). *Optimal transport: old and new*, volume 338. Springer.

Wager, S. (2020). Stats 361: Causal inference.

Wang, W., Janson, L., Lei, L., and Ramdas, A. (2023). Total variation floodgate for variable importance inference in classification.

Zaev, D. A. (2015). On the monge–kantorovich problem with additional linear constraints. *Mathematical Notes*, 98:725–741.

Zhang, J. L. and Rubin, D. B. (2003). Estimation of causal effects via principal stratification when some outcomes are truncated by "death". *Journal of Educational and Behavioral Statistics*, 28(4):353–368.

Zhang, L. and Janson, L. (2020). Floodgate: inference for model-free variable importance.

Zrnic, T. and Fithian, W. (2022). Locally simultaneous inference.

Zualinescu, C. (2003). Sharp estimates for hoffman's constant for systems of linear inequalities and equalities. *SIAM Journal on Optimization*, 14(2):517–533.

A Additional methodological details

A.1 Choosing the minimum norm solution when $\hat{\nu}$ is not unique

Our suggested strategy to compute optimal dual variables involves solving the following optimization problem over $\nu_{0,x}, \nu_{1,x} : \mathcal{Y} \to \mathbb{R}$:

$$\hat{\nu}_{0,x}, \hat{\nu}_{1,x} = \underset{\nu_{0,x},\nu_{1,x} \in \mathcal{V}_x}{\arg\max} \mathbb{E}_{\hat{P}_{Y(0)|X=x}}[\nu_{0,x}(Y(0))] + \mathbb{E}_{\hat{P}_{Y(1)|X=x}}[\nu_{1,x}(Y(1))]. \tag{40}$$

This problem does not always have a unique solution. For most of our theory, this does not matter; the theorems will hold if one computes any solution to this equation. However, practically speaking, it may be helpful to pick the minimum norm solution to reduce the variance of the final estimator. Furthermore, Lemma 3.3 specifically assumes that we take the minimum norm solution (as proposed in Section 2). In this section, we formalize the notion of the minimum norm solution and discuss how to compute it.

Precisely, we suggest taking the minimum norm solution with respect to the L^2 inner product on $\mathcal{V}_x$. In particular, assume that $Y(0) \mid X = x$ and $Y(1) \mid X = x$ have conditional densities with respect to some base measure ψ_x on $\mathcal{Y}$. (E.g., we choose ψ_x to be the Lebesgue measure for continuous potential outcomes and the counting measure for discrete potential outcomes.) Then the inner product is defined as

$$\langle (\nu_{0,x}, \nu_{1,x}), (\nu'_{0,x}, \nu'_{1,x}) \rangle \coloneqq \int \nu_{0,x}(y_0)\nu'_{0,x}(y_0)\psi_x(dy_0) + \int \nu_{1,x}(y_1)\nu'_{1,x}(y_1)\psi_x(dy_1). \tag{41}$$

We note that in some settings, all solutions to Eq. (40) may have infinite norms. In this case, we recommend just picking a solution at random, since any solution is a minimum norm solution.

To compute the minimum norm solution, we recommend using the discretization scheme from Section 4. In particular, we approximate $Y(0), Y(1)$ as discrete variables on finite sets $\mathcal{Y}_0 = \{y_{0,1}, \ldots, y_{0,n_{\text{vals}},x}\}, \mathcal{Y}_1 = \{y_{1,1}, \ldots, y_{1,n_{\text{vals}},x}\}$ with conditional PMFs $\{p_{0,j,x}\}_{j=1}^{n_{\text{vals}}}, \{p_{1,i,x}\}_{i=1}^{n_{\text{vals}}}$. Then, we can approximately solve Eq. (40) by solving the following linear program:

$$\begin{aligned}
&\max \sum_{j=1}^{n_{\text{vals}}} p_{0,j,x}\nu_{0,x}(y_{0,j,x}) + \sum_{i=1}^{n_{\text{vals}}} p_{1,i,x}\nu_{1,x}(y_{1,i,x}) \\
&\text{s.t. } \nu_{0,x}(y_{0,j,x}) + \nu_{1,x}(y_{1,i,x}) - \sum_{\ell=1}^{L} \lambda_{x,\ell} w_{x,\ell}(y_{0,j,x}, y_{1,i,x}) \leq f(y_{0,j,x}, y_{1,i,x}, x) \text{ for all } i, j \in [n_{\text{vals}}] \\
&\quad \lambda_{x,1}, \ldots, \lambda_{x,L} \geq 0.
\end{aligned}$$

To find an (approximate) minimum norm solution, we first solve the original version of this linear program and find the optimal objective value $\hat{o}$ for the linear program. Then, to find a minimum norm solution, we solve the new convex quadratic program which minimizes the norm over all optimal solutions:

$$\begin{aligned}
&\min \sum_{j=1}^{n_{\text{vals}}} \nu_{0,x}(y_{0,j,x})^2 + \sum_{i=1}^{n_{\text{vals}}} \nu_{1,x}(y_{1,i,x})^2 \\
&\text{s.t. } \sum_{j=1}^{n_{\text{vals}}} p_{0,j,x}\nu_{0,x}(y_{0,j,x}) + \sum_{i=1}^{n_{\text{vals}}} p_{1,i,x}\nu_{1,x}(y_{1,i,x}) = \hat{o} \\
&\quad \nu_{0,x}(y_{0,j,x}) + \nu_{1,x}(y_{1,i,x}) - \sum_{\ell=1}^{L} \lambda_{x,\ell} w_{x,\ell}(y_{0,j,x}, y_{1,i,x}) \leq f(y_{0,j,x}, y_{1,i,x}, x) \text{ for all } i, j \in [n_{\text{vals}}] \\
&\quad \lambda_{x,1}, \ldots, \lambda_{x,L} \geq 0.
\end{aligned}$$

After solving this convex quadratic program, one can obtain full estimated dual variables $\hat{\nu}_{0,x}, \hat{\nu}_{1,x}$ using the interpolation and grid-search scheme introduced in Section 4.1 and 4.2.

A.2 Inference and model selection for generalized estimands with cross-fitting

This paper primarily considers partially identifiable estimands of the form $\theta(P^\star) = \mathbb{E}_{P^\star}[f(Y(1), Y(0), X)]$. However, many estimands can be written in the form

$$\theta(P^\star) = h\Big(\mathbb{E}_{P^\star}[f(Y(1), Y(0), X)],\ \mathbb{E}_{P^\star}[z_1(Y(1), X)],\ \mathbb{E}_{P^\star}[z_0(Y(0), X)]\Big). \tag{42}$$

for some functions $z_0 : \mathcal{Y} \times \mathcal{X} \to \mathbb{R}^{d_0}, z_1 : \mathcal{Y} \times \mathcal{X} \to \mathbb{R}^{d_1}$ and $h : \mathbb{R}^{d_0+d_1+1} \to \mathbb{R}$ such that h is nondecreasing in its first argument and is continuously differentiable. In other words, $\theta(P^\star)$ can be written as a (nonlinear) function of a partially identifiable expectation and two identifiable expectations. We give two examples of this below.

Example 9 (Variance of the ITE)**.** If $\theta(P^\star) = \text{Var}(Y(1) - Y(0))$, we can write

$$\theta(P^\star) = \mathbb{E}_{P^\star}[(Y(1) - Y(0))^2] - (\mathbb{E}[Y(1)] - \mathbb{E}[Y(0)])^2 \tag{43}$$

which satisfies Eq. (42) if we set $f(y_1, y_0, x) = (y_1 - y_0)^2$, $z_0(y_0, x) = y_0$, $z_1(y_1, x) = y_1$, and $h(a, b, c) = a - (b - c)^2$.

Example 10 (Lee bounds under monotonicity)**.** In the case of Lee bounds (Ex 3) under monotonicity, we have compound potential outcomes of the form $Y(0), S(0)$ and $Y(1), S(1)$ and the estimand can be written as

$$\theta(P^\star) = \frac{\mathbb{E}_{P^\star}\left[(Y(1) - Y(0))\, S(0)\right]}{\mathbb{E}_{P^\star}[S(0)]}$$

which satisfies Eq. (42) if we set $f((y_1, s_1), (y_0, s_0), x) = (y_1 - y_0)s_0$, $z_0((y_0, s_0), x) = s_0$, $z_1((y_1, s_1), x) = 0$, and $h(a, b, c) = a/c$.

We now show how to perform inference on estimands in the general case of Eq. (42). First, we give the main idea without discussing cross-fitting or model selection. Then, we introduce a multiplier bootstrap-like method to select the tightest bounds among K cross-fit estimators of $\theta(P^\star)$.

A.2.1 Main idea

As notation, let $\beta = \mathbb{E}_{P^\star}[f(Y(1), Y(0), X)] \in \mathbb{R}$, $\kappa_1 = \mathbb{E}_{P^\star}[z_1(Y(1), X)] \in \mathbb{R}^{d_1}$ and $\kappa_0 = \mathbb{E}_{P^\star}[z_0(Y(0), X)] \in \mathbb{R}^{d_0}$ so that $\theta(P^\star) = h(\beta, \kappa_1, \kappa_0)$. If β_L is the sharp lower bound on β, then $\theta_L \coloneqq h(\beta_L, \kappa_1, \kappa_0)$ is the sharp lower bound on $\theta(P^\star)$ since h is monotone in its first coordinate and κ_1, κ_0 are identified.

The main idea is as follows. For the partially identified term $\mathbb{E}_{P^\star}[f(Y(1), Y(0), X)]$, estimate dual variables $\hat{\nu}_0, \hat{\nu}_1 : \mathcal{Y} \times \mathcal{X} \to \mathbb{R}$ from $\mathcal{D}_1$ using techniques from the rest of the paper such that weak duality holds, that is, $\mathbb{E}_P[\hat{\nu}_0(Y(0), X) + \hat{\nu}_1(Y(1), X) \mid \mathcal{D}_1] \leq \mathbb{E}_P[f(Y(1), Y(0), X)]$ for all $P \in \mathcal{P}$. Then define the IPW estimators

$$\hat{\beta} = \frac{1}{|\mathcal{D}_2|} \sum_{i \in \mathcal{D}_2} S_i^{(\beta)} \text{ for } S_i^{(\beta)} \coloneqq \frac{W_i \hat{\nu}_1(Y_i, X_i)}{\pi(X_i)} + \frac{(1 - W_i)\hat{\nu}_0(Y_i, X_i)}{1 - \pi(X_i)} \tag{44}$$

and for $w \in \{0, 1\}$,

$$\hat{\kappa}_w = \frac{1}{|\mathcal{D}_2|} \sum_{i \in \mathcal{D}_2} S_i^{(\kappa_w)} \text{ for } S_i^{(\kappa_w)} \coloneqq \frac{\mathbb{I}(W_i = w) z_w(Y_i, X_i)}{w\pi(X_i) + (1 - w)(1 - \pi(X_i))}. \tag{45}$$

The multivariate CLT says that under appropriate moment conditions, conditional on $\mathcal{D}_1$ we have that

$$\sqrt{|\mathcal{D}_2|} \left(\begin{bmatrix} \hat{\beta} \\ \hat{\kappa}_1 \\ \hat{\kappa}_0 \end{bmatrix} - \begin{bmatrix} \tilde{\beta} \\ \kappa_1 \\ \kappa_0 \end{bmatrix} \right) \xrightarrow{\mathrm{d}} \mathcal{N}(0, \Sigma) \tag{46}$$

where $\tilde{\beta} = \mathbb{E}[\hat{\beta} \mid \mathcal{D}_1] \leq \beta$ by weak duality and $\Sigma \coloneqq \mathrm{Cov}\left(\left(S_i^{(\beta)}, S_i^{(\kappa_1)}, S_i^{(\kappa_0)} \right) \mid \mathcal{D}_1 \right)$. The delta method yields that

$$\sqrt{|\mathcal{D}_2|} \left(h\left(\hat{\beta}, \hat{\kappa}_1, \hat{\kappa}_0 \right) - h\left(\tilde{\beta}, \kappa_1, \kappa_0 \right) \right) \xrightarrow{\mathrm{d}} \mathcal{N}\left(0, \nabla h\left(\tilde{\beta}, \kappa_1, \kappa_0 \right)^T \Sigma \nabla h\left(\tilde{\beta}, \kappa_1, \kappa_0 \right) \right).$$

By plugging in $\hat{\beta}, \hat{\kappa}_1, \hat{\kappa}_0$, we can get a consistent estimator of the gradient $\nabla h\left(\tilde{\beta}, \kappa_1, \kappa_0 \right)$. Furthermore, we can get a consistent estimator of Σ by letting $\hat{\Sigma}$ denote the empirical covariance matrix of the conditionally i.i.d. vectors $\{(S_i^{(\beta)}, S_i^{(\kappa_1)}, S_i^{(\kappa_0)})\}_{i \in \mathcal{D}_2}$. If we set $\hat{\theta}_L = h\left(\hat{\beta}, \hat{\kappa}_1, \hat{\kappa}_0 \right)$ and $\tilde{\theta}_L = h\left(\tilde{\beta}, \kappa_1, \kappa_0 \right)$, Slutsky's theorem yields

$$\sqrt{\frac{|\mathcal{D}_2|}{\nabla h\left(\hat{\beta}, \hat{\kappa}_1, \hat{\kappa}_0 \right)^T \hat{\Sigma} \nabla h\left(\hat{\beta}, \hat{\kappa}_1, \hat{\kappa}_0 \right)}} \left(\hat{\theta}_L - \tilde{\theta}_L \right) \xrightarrow{\mathrm{d}} \mathcal{N}(0, 1).$$

Using this equation, we note that

$$\hat{\theta}_{\mathrm{LCB}} = \hat{\theta}_L - \Phi^{-1}(1 - \alpha) \sqrt{\frac{\nabla h\left(\hat{\beta}, \hat{\kappa}_1, \hat{\kappa}_0 \right)^T \hat{\Sigma} \nabla h\left(\hat{\beta}, \hat{\kappa}_1, \hat{\kappa}_0 \right)}{|\mathcal{D}_2|}}$$

is an asymptotic $1 - \alpha$ lower confidence bound on $\tilde{\theta}_L$. Note that by weak duality, $\tilde{\beta} \leq \beta$, and therefore since h is nondecreasing in its first argument, we have that

$$\tilde{\theta}_L = h(\tilde{\beta}, \kappa_1, \kappa_0) \leq h(\beta, \kappa_1, \kappa_0) = \theta(P^\star)$$

and therefore $\hat{\theta}_L$ is a valid lower confidence bound on $\theta(P^\star)$ as well.

Remark 10. This calculation requires that the dimensions d_0, d_1 are fixed constants that do not grow with n.

A.2.2 Cross-fitting and model-selection

In Section 2.3, we introduced a multiplier bootstrap method that selects the tightest possible dual bounds across K dual variable estimates (e.g., fit using different subsets of the covariates), where K may grow exponentially with n. We now generalize this method in two ways. First, we now permit the use of cross-fitting. Second, we consider the generalized class of estimands defined in Eq. (42). However, this generality

comes at a cost: unlike Corollary 3.1, we require that the number of dual variable estimates K is fixed and does not grow with n.

We first define the method; then we prove its validity. Suppose given the first fold of data $\mathcal{D}_1$, we produce K candidate dual variables $\hat{\nu}^{(1)}, \ldots, \hat{\nu}^{(K)} \in \mathcal{V}$, and symmetrically using the second fold $\mathcal{D}_2$ we produce $\hat{\nu}^{(1,\text{swap})}, \ldots, \hat{\nu}^{(K,\text{swap})} \in \mathcal{V}$. For ease of exposition, we assume n is even and $|\mathcal{D}_1| = |\mathcal{D}_2| = n/2$. Again, the results in this section can be easily extended to M-fold cross-fitting for $M > 2$. For each $k \in [K]$, the cross-fit dual lower estimate of $\theta(P^\star)$ is defined by plugging in an IPW-mean estimator of κ_1, κ_0 and a dual cross-fit lower estimator of $\beta \coloneqq \mathbb{E}_{P^\star}[f(Y(1), Y(0), X)]$ into the definition $\theta(P^\star) = h(\beta, \kappa_1, \kappa_0)$. Precisely:

$$\hat{\theta}_L^{(k)} \coloneqq h(\hat{\beta}^{(k)}, \hat{\kappa}_1, \hat{\kappa}_0), \tag{47}$$

where

$$\hat{\beta}^{(k)} = \frac{1}{n}\sum_{i=1}^{n} S_i^{(\beta,k)} \text{ for } S_i^{(\beta,k)} \coloneqq \begin{cases} \frac{W_i \hat{\nu}_1^{(k)}(Y_i,X_i)}{\pi(X_i)} + \frac{(1-W_i)\hat{\nu}_0^{(k)}(Y_i,X_i)}{1-\pi(X_i)} & i \in \mathcal{D}_2 \\ \frac{W_i \hat{\nu}_1^{(k,\text{swap})}(Y_i,X_i)}{\pi(X_i)} + \frac{(1-W_i)\hat{\nu}_0^{(k,\text{swap})}(Y_i,X_i)}{1-\pi(X_i)} & i \in \mathcal{D}_1, \end{cases} \tag{48}$$

and for $w \in \{0, 1\}$,

$$\hat{\kappa}_w = \frac{1}{n}\sum_{i=1}^{n} S_i^{(\kappa_w)} \text{ for } S_i^{(\kappa_w)} \coloneqq \frac{\mathbb{I}(W_i = w) z_w(Y_i, X_i)}{w\pi(X_i) + (1-w)(1-\pi(X_i))}. \tag{49}$$

The standard error $\hat{\sigma}^{(k)}$ of $\sqrt{n}\hat{\theta}_L^{(k)}$ is defined as:

$$\hat{\sigma}^{(k)} = \sqrt{\nabla h\left(\hat{\beta}^{(k)}, \hat{\kappa}_1, \hat{\kappa}_0\right)^T \hat{\Sigma}^{(k)} \nabla h\left(\hat{\beta}^{(k)}, \hat{\kappa}_1, \hat{\kappa}_0\right)}, \tag{50}$$

where $\hat{\Sigma}^{(k)} \in \mathbb{R}^{(1+d_1+d_0)\times(1+d_1+d_0)}$ is the empirical covariance matrix of the vectors $(S_i^{(\beta,k)}, S_i^{(\kappa_1)}, S_i^{(\kappa_0)})$ for $i \in [n]$. To aggregate evidence across all K lower confidence bounds, we require the following notation. Let $\hat{\Sigma}_{\text{full}} \in \mathbb{R}^{(K+d_0+d_1)\times(K+d_0+d_1)}$ denote the empirical covariance matrix of $\vec{S}_i = (S_i^{(\beta,1)}, \ldots, S_i^{(\beta,K)}, S_i^{(\kappa_1)}, S_i^{(\kappa_0)}) \in \mathbb{R}^{K+d_0+d_1}$ and let $H : \mathbb{R}^{K+d_0+d_1} \to \mathbb{R}^K$ be the function defined by $H_k(x) = h(x_k, x_{(K+1):(K+d_0+d_1)})$. In particular, this definition ensures that if $\bar{S}$ is the sample average of $\{\vec{S}_i\}_{i\in[n]}$, then $H(\bar{S}) = (\hat{\theta}_L^{(1)}, \ldots, \hat{\theta}_L^{(K)})$. Define

$$\hat{\Sigma}_H = \nabla H(\bar{S})^T \hat{\Sigma}_{\text{full}} \nabla H(\bar{S}) \text{ and } \hat{C}_H = \text{diag}\left\{\hat{\Sigma}_H\right\}^{-1/2} \hat{\Sigma}_H \text{diag}\left\{\hat{\Sigma}_H\right\}^{-1/2}. \tag{51}$$

Then the final combined lower bound is defined as

$$\hat{\theta}_{\text{LCB}}^{\text{crossfit}} = \max_{k=1}^{K} \hat{\theta}_L^{(k)} - \hat{q}_{1-\alpha}\frac{\hat{\sigma}^{(k)}}{\sqrt{n}}, \tag{52}$$

where we define $\hat{q}_{1-\alpha}$ as the $1-\alpha$ quantile of the maximum of a $\mathcal{N}(0, \hat{C}_H)$ vector:

$$\hat{q}_{1-\alpha} \coloneqq Q_{1-\alpha}\left(\max_{k=1}^{K} Z_k\right) \text{ for } Z \sim \mathcal{N}(0, \hat{C}_H). \tag{53}$$

We now show that Eq. (52) defines a valid lower confidence bound under essentially the same assumptions as Proposition 3.2 as long as the number of models K does not grow with n. (We implicitly assume that the functions defining the estimand—namely h, f, z_0, z_1—do not change with n.) Below, note that for dual variables $\nu \in \mathcal{V}$, $g(\nu) = \sum_{w\in\{0,1\}} \mathbb{E}_{P^\star}[\nu_w(Y(w), X)]$ is the Lagrange dual function from Section 2.1.

Corollary A.1. *Suppose that h is continuously differentiable and nondecreasing in its first argument. Under Assumption 3.1, for $\alpha \leq 0.5$,*

$$\liminf_{n\to\infty} \mathbb{P}(\hat{\theta}_{\text{LCB}}^{\text{crossfit}} \leq \theta_L) \geq 1 - \alpha,$$

holds as long as for each $k \in [K]$, $\hat{\nu}^{(k)}$ satisfies Assumption 3.2 and one of the two following conditions:

1. *Condition 1: There exist arbitrary deterministic dual variables $\nu^{(k,\dagger)} \in \mathcal{V}$ satisfying Assumption 3.2 such that $\mathbb{E}\left[\left(\hat{\nu}_w^{(k)}(Y(k), X) - \nu_w^{(k,\dagger)}(Y(k), X)\right)^2\right] \to 0$ holds at any rate for $w \in \{0, 1\}$. Note that we do not allow $\{\nu_w^\dagger\}_{k\in\{0,1\}}$ to change with n. Furthermore, if $S_i^{(\beta,k,\dagger)}$ is defined analogously to $S_i^{(\beta,k)}$ but with $\nu^{(k,\dagger)}$ replacing $\hat{\nu}^{(k)}$ and $\hat{\nu}^{(k,\text{swap})}$, then we require that*

$$\nabla h\left(g(\nu^{(k,\dagger)}), \kappa_1, \kappa_0\right)^T \text{Cov}(S_i^{(\beta,k,\dagger)}, S_i^{\kappa_1}, S_i^{\kappa_0}) \nabla h\left(g(\nu^{(k,\dagger)}), \kappa_1, \kappa_0\right) > 0. \tag{54}$$

2. *Condition 2: The outcome model is sufficiently misspecified such that the first-stage bias is larger than* $n^{-1/2}$*, i.e.,* $n^{-1/2}\left(\beta_L - \frac{g(\hat{\nu}^{(k)})+g(\hat{\nu}^{(k,\mathrm{swap})})}{2}\right) \xrightarrow{\mathrm{p}} \infty$*. Furthermore, the partial derivative* $\partial_b h(b, \kappa_1, \kappa_0)$ *is bounded away from zero for all* $b \in \mathbb{R}$*.*

Remark 11. We recommend that the reader read the proofs of Theorem 3.1 and Proposition 3.2 before reading this proof.

Remark 12. Condition 1 and Condition 2 are the same conditions required in Proposition 3.2, with three changes. First, for simplicity, we do not allow $\nu^{(k,\dagger)}$ to change with n. Second, we require the condition Eq. (54), which ensures that the limiting variance of $\sqrt{n}\hat{\theta}_L^{(k)}$, as calculated by the delta method, is nonzero. Note that a similar "nonzero variance" condition already appears in Proposition 3.2 via Assumption 3.2. Third, in Condition 2, we require a lower bound on the partial derivative of h with respect to its first coordinate. This is necessary to guarantee that if $\hat{\beta}^{(k)}$ is asymptotically conservative for β_L, then $\hat{\theta}_L^{(k)}$ will be conservative for θ_L.

Proof. We handle the two conditions separately.

Condition 1: We first prove the result in the special case where $\hat{\nu}^{(k)}$ satisfies Condition 1 for every $k \in [K]$. As notation, let $S_i^{(k,\beta,\dagger)}, \hat{\beta}^{(k,\dagger)}, \hat{\sigma}^{(k,\dagger)}, \hat{\Sigma}_{\mathrm{full}}^{\dagger}, \vec{S}_i^{\dagger}$ be defined analogously to $S_i^{(k,\beta)}, \hat{\beta}^{(k)}, \hat{\sigma}^{(k)}, \hat{\Sigma}_{\mathrm{full}}, \vec{S}_i$ but replacing $\hat{\nu}^{(k)}$ with $\nu^{(k,\dagger)}$, for each $k \in [K]$. The proof of Proposition 3.2 in Appendix D.6 shows the following relationships between these quantities:

1. $\hat{\beta}^{(k)} \le \hat{\beta}^{(k,\dagger)} + \Delta_k + o_p(n^{-1/2})$, where (a) $\Delta_k \le \beta_L - \mathbb{E}[\hat{\beta}^{(k,\dagger)}]$ and (b) $\Delta_k = o_p(1)$.
2. $\hat{\sigma}^{(k)} - \hat{\sigma}^{(k,\dagger)} = o_p(1)$, and a similar argument shows $\hat{\Sigma}_{\mathrm{full}} - \hat{\Sigma}_{\mathrm{full}}^{\dagger} = o_p(1)$ holds elementwise (this follows from a uniform law of large numbers as reviewed in Appendix D.6). Note that $\hat{\Sigma}_{\mathrm{full}}^{\dagger}$ is simply an empirical covariance matrix of the i.i.d. vectors $\vec{S}_i^{\dagger} \coloneqq (S_i^{(\beta,1,\dagger)}, \ldots, S_i^{(\beta,K,\dagger)}, S_i^{(\kappa_1)}, S_i^{(\kappa_0)})$, for $i \in [n]$. Thus, this also implies that $\hat{\Sigma}_{\mathrm{full}}, \hat{\Sigma}_{\mathrm{full}}^{\dagger} \xrightarrow{\mathrm{p}} \Sigma_{\mathrm{full}}^{\dagger} \coloneqq \mathrm{Cov}(\vec{S}_i^{\dagger})$.

These results imply the following results:

3. The first result implies that

$$\hat{\theta}_L^{(k)} \coloneqq h(\hat{\beta}^{(k)}, \hat{\kappa}_1, \hat{\kappa}_0) \le h(\hat{\beta}^{(k,\dagger)} + \Delta_k + o_p(n^{-1/2}), \hat{\kappa}_1, \hat{\kappa}_0) \le h(\hat{\beta}^{(k,\dagger)} + \Delta_k, \hat{\kappa}_1, \hat{\kappa}_0) + o_p(n^{-1/2}).$$

   The first inequality follows because h is nondecreasing in its first argument. The second argument follows because h is continuously differentiable and $\hat{\beta}^{(k,\dagger)} + \Delta_k, \hat{\kappa}_1, \hat{\kappa}_0$ converges uniformly to $(\mathbb{E}[\hat{\beta}^{(k,\dagger)}], \kappa_1, \kappa_0)$ by the law of large numbers (remember that all quantities involved have bounded $2+\delta$ moments by assumption). Thus, h is locally Lipschitz at $(\mathbb{E}[\hat{\beta}^{(k,\dagger)}], \kappa_1, \kappa_0)$ and the result holds.
4. Define $C_H^{\dagger} \coloneqq \mathrm{diag}\left\{\Sigma_H^{\dagger}\right\}^{-1/2} \Sigma_H^{\dagger} \mathrm{diag}\left\{\Sigma_H^{\dagger}\right\}^{-1/2}$ where $\Sigma_H^{\dagger} \coloneqq \nabla H(\mathbb{E}[\vec{S}_i^{\dagger}])^T \Sigma_{\mathrm{full}}^{\dagger} \nabla H(\mathbb{E}[\vec{S}_i^{\dagger}])$. In words, $C_H^{\dagger}$ is essentially the population variant of $\hat{C}_H$. Since $\bar{S} \xrightarrow{\mathrm{p}} \mathbb{E}[S_i^{\dagger}]$ and $\hat{\Sigma}_{\mathrm{full}} \xrightarrow{\mathrm{p}} \Sigma_{\mathrm{full}}^{\dagger}$ and ∇H is continuous by assumption, we know that $\hat{C}_H \xrightarrow{\mathrm{p}} C_H^{\dagger}$. Thus, the continuous mapping theorem yields

$$\hat{q}_{1-\alpha} \xrightarrow{\mathrm{p}} q_{1-\alpha}^{\dagger} \coloneqq Q_{1-\alpha}\left(\max_{k=1}^{K} Z_k\right) \text{ for } Z \sim \mathcal{N}(0, C_H). \tag{55}$$

Since K does not grow with n, we can combine the second, third, and fourth results to obtain:

$$\hat{\theta}_{\mathrm{LCB}}^{\mathrm{crossfit}} \coloneqq \max_{k=1}^{K} \hat{\theta}_L^{(k)} - \hat{q}_{1-\alpha}\frac{\hat{\sigma}^{(k)}}{\sqrt{n}} \le \max_{k=1}^{K} h(\hat{\beta}^{(k,\dagger)} + \Delta_k, \hat{\kappa}_1, \hat{\kappa}_0) - q_{1-\alpha}^{\dagger}\frac{\hat{\sigma}^{(k,\dagger)}}{\sqrt{n}} + o_p(n^{-1/2}). \tag{56}$$

As notation, let $\hat{\kappa} = (\hat{\kappa}_1, \hat{\kappa}_0) \in \mathbb{R}^{d_1+d_0}$ and $\kappa = [\kappa_1, \kappa_0] \in \mathbb{R}^{d_1+d_0}$. Then observe

$$\begin{aligned}
\star &\coloneqq \mathbb{P}\left(\max_{k=1}^{K} \hat{\theta}_L^{(k)} - \hat{q}_{1-\alpha}\frac{\hat{\sigma}^{(k)}}{\sqrt{n}} \le \theta_L\right) \\
&= \mathbb{P}\left(\max_{k=1}^{K} h(\hat{\beta}^{(k,\dagger)} + \Delta_k, \hat{\kappa}) - q_{1-\alpha}^{\dagger}\frac{\hat{\sigma}^{(k,\dagger)}}{\sqrt{n}} + o_p(n^{-1/2}) \le \theta_L\right) && \text{by Eq. (56)} \\
&= \mathbb{P}\left(\max_{k=1}^{K} \frac{\sqrt{n}\left(h(\hat{\beta}^{(k,\dagger)} + \Delta_k, \hat{\kappa}) - \theta_L\right)}{\hat{\sigma}^{(k,\dagger)}} + o_p(1) \le q_{1-\alpha}^{\dagger}\right) && \text{by rearrangement.}
\end{aligned}$$

Now, we have essentially replaced $\hat{\nu}^{(k)}$ with $\hat{\nu}^{(k,\dagger)}$ for each k—the next step is to eliminate the random (and non-negligible) Δ_k. We will do this by replacing each Δ_k with a constant a_k; later, we will let $a_k \to 0$.

To be precise, recall that for each k, $\Delta_k \leq \beta_L - \mathbb{E}[\hat{\beta}^{(k,\dagger)}]$ *and* $\Delta_k = o_p(1)$. Thus, for each k, we may pick a constant $a_k \geq 0$ such that (i) $\Delta_k \leq a_k$ with probability approaching one asymptotically[9] and (ii) $a_k \leq \beta_L - \mathbb{E}[\hat{\beta}^{(k,\dagger)}]$. Since h is nondecreasing in its first argument, this implies that (i) $h(\hat{\beta}^{(k,\dagger)} + \Delta_k, \hat{\kappa}) \leq h(\hat{\beta}^{(k,\dagger)} + a_k, \hat{\kappa})$ holds asymptotically with probability approaching one and (ii) $\theta_L = h(\beta_L, \kappa) \geq h(\mathbb{E}[\hat{\beta}^{(k,\dagger)}] + a_k, \kappa)$. Thus, since K is finite, asymptotically we have that:

$$\liminf_{n\to\infty} \star \geq \liminf_{n\to\infty} \mathbb{P}\left(\max_{k=1}^{K} \frac{\sqrt{n}\left(h(\hat{\beta}^{(k,\dagger)} + a_k, \hat{\kappa}) - h(\mathbb{E}[\hat{\beta}^{(k,\dagger)}] + a_k, \kappa)\right)}{\hat{\sigma}^{(k,\dagger)}} + o_p(1) \leq q^{\dagger}_{1-\alpha} \right).$$

Now, we have successfully replaced the Δ_k's with constants a_k's. Our next step is to modify the denominator $\hat{\sigma}^{(k,\dagger)}$ and replace it with one that accounts for the influence of a_k, and then bound the error from this approximation. As notation, let $\hat{\sigma}^{(k,a_k,\dagger)}$ be defined analogously to $\sigma^{(k,\dagger)}$ but replacing $\hat{\beta}^{(k,\dagger)}$ with $\hat{\beta}^{(k,\dagger)} + a_k$, that is,

$$\hat{\sigma}^{(k,a_k,\dagger)} = \sqrt{\nabla h\left(\hat{\beta}^{(k,\dagger)} + a_k, \hat{\kappa}\right)^T \hat{\Sigma}^{(k,\dagger)} \nabla h\left(\hat{\beta}^{(k,\dagger)} + a_k, \hat{\kappa}\right)}$$

and let $\sigma^{(k,a_k,\dagger)}$ be the population variant:

$$\sigma^{(k,a_k,\dagger)} = \sqrt{\nabla h\left(\mathbb{E}[\hat{\beta}^{(k,\dagger)}] + a_k, \kappa\right)^T \Sigma^{(k,\dagger)} \nabla h\left(\mathbb{E}[\hat{\beta}^{(k,\dagger)}] + a_k, \kappa\right)}.$$

Lastly, let $\hat{Z}_k = \frac{\sqrt{n}}{\sigma^{(k,a_k,\dagger)}}(h(\hat{\beta}^{(k,\dagger)} + a_k, \hat{\kappa}) - h(\mathbb{E}[\hat{\beta}^{(k,\dagger)}] + a_k, \kappa))$. Rearranging, we obtain

$$\liminf_{n\to\infty} \star \geq \liminf_{n\to\infty} \mathbb{P}\left(\max_{k=1}^{K} \frac{\sigma^{(k,a_k,\dagger)}}{\hat{\sigma}^{(k,\dagger)}} \hat{Z}_k + o_p(1) \leq q^{\dagger}_{1-\alpha} \right) \tag{57}$$

$$\geq \liminf_{n\to\infty} \mathbb{P}\left(\max_{k=1}^{K} \hat{Z}_k + \max_{k=1}^{K} \left(\frac{\sigma^{(k,a_k,\dagger)}}{\hat{\sigma}^{(k,\dagger)}} - 1 \right) \hat{Z}_k + o_p(1) \leq q^{\dagger}_{1-\alpha} \right). \tag{58}$$

Now, we observe that $\hat{Z} \coloneqq (\hat{Z}_1, \ldots, \hat{Z}_K)$ is asymptotically multivariate Gaussian by the multivariate delta method. In particular, define the vector of summands

$$V_i \coloneqq (S_i^{(\beta,1,\dagger)} + a_1, \ldots, S_i^{(\beta,K,\dagger)} + a_K, S_i^{(\kappa_1)}, S_i^{(\kappa_0)}) \in \mathbb{R}^{K+d_0+d_1}$$

and let $\bar{V} = \frac{1}{n}\sum_{i=1}^{n} V_i$. If we define $\tilde{\beta}_L = (\mathbb{E}[\hat{\beta}^{(1,\dagger)}], \ldots, \mathbb{E}[\hat{\beta}^{(K,\dagger)}])$ and $\vec{a} = (a_1, \ldots, a_K)$, the multivariate CLT yields that

$$\sqrt{n}(\bar{V} - (\tilde{\beta}_L + \vec{a}, \kappa)) \xrightarrow{\mathrm{d}} \mathcal{N}(0, \Sigma_{\text{full}}), \tag{59}$$

where notably Σ_{full} does not depend on $\vec{a}$. For the continuously differentiable function $H : \mathbb{R}^{K+d_0+d_1} \to \mathbb{R}^K$ defined by $H_k(\bar{V}) \coloneqq h(\bar{V}_k, \bar{V}_{K+1:(K+d_0+d_1)})) = h(\hat{\beta}^{(k,\dagger)} + a_k, \hat{\kappa})$, the multivariate delta method yields

$$\hat{Z} \xrightarrow{\mathrm{d}} \mathcal{N}(0, C_{H,\vec{a}}) \text{ where } C_{H,\vec{a}} \text{ is the correlation matrix of } \Sigma_{H,\vec{a}} \coloneqq \nabla H(\tilde{\beta}_L + \vec{a}, \kappa)^T \Sigma_{\text{full}} \nabla H(\tilde{\beta}_L + \vec{a}, \kappa).$$

(Note that $\Sigma_{H,\vec{a}}$ has nonzero diagonal entries for all $\vec{a}$ sufficiently close to zero because $\Sigma_{H,\vec{0}} = \Sigma_H$ has nonzero diagonal entries by assumption and ∇H is assumed to be continuous.) Thus, by the continuous mapping theorem, we conclude that as $n \to \infty$,

$$\max_{k=1}^{K} \hat{Z}_k + \max_{k=1}^{K} \left(\frac{\sigma^{(k,a_k,\dagger)}}{\hat{\sigma}^{(k,\dagger)}} - 1 \right) \hat{Z}_k \xrightarrow{\mathrm{d}} \max_{k=1}^{K} Z_k + \max_{k=1}^{K} \left(\frac{\sigma^{(k,a_k,\dagger)}}{\sigma^{(k,\dagger)}} - 1 \right) Z_k \text{ for } Z \sim \mathcal{N}(0, C_{H,\vec{a}}).$$

This implies

$$\liminf_{n\to\infty} \star \geq \mathbb{P}_{Z\sim\mathcal{N}(0,C_{H,\vec{a}})} \left(\max_{k=1}^{K} Z_k + \max_{k=1}^{K} \left(\frac{\sigma^{(k,a_k,\dagger)}}{\sigma^{(k,\dagger)}} - 1 \right) Z_k \leq q^{\dagger}_{1-\alpha} \right),$$

where this holds for all $\vec{a}$ sufficiently close to zero. Note that by assumption, the gradient of h is continuous; thus ∇H is continuous as well. Thus, taking the limit as $\vec{a} \to 0$, we obtain

$$\liminf_{n\to\infty} \star \geq \mathbb{P}_{Z\sim\mathcal{N}(0,C_{H,\vec{0}})} \left(\max_{k=1}^{K} Z_k \leq q^{\dagger}_{1-\alpha} \right) = 1 - \alpha,$$

[9] That is, $\mathbb{I}(\Delta_k \leq a_k) \xrightarrow{\mathrm{p}} 1$, although the convergence does not necessarily hold a.s.

where the right-hand equality holds by definition of $\hat{q}_{1-\alpha}$.

Condition 2: We now consider the general case. Without loss of generality, suppose $k = 1, \ldots, K_0$ satisfy Condition 1, and $k = K_0 + 1, \ldots, K$ satisfy Condition 2. Note that the proof for Condition 1 shows that

$$\liminf_{n\to\infty} \mathbb{P}\left(\max_{k=1}^{K_0} \hat{\theta}_L^{(k)} - \hat{q}_{1-\alpha} \frac{\hat{\sigma}^{(k)}}{\sqrt{n}} \leq \theta_L \right) \geq 1 - \alpha,$$

where in particular this holds because the addition of $\hat{\nu}^{(K_0+1)}, \ldots, \hat{\nu}^{(K)}$ does not affect the values of $\hat{\theta}_L^{(k)}, \hat{\sigma}^{(k)}$ and can only increase the value of $\hat{q}_{1-\alpha}$. Thus, it suffices to show that

$$\liminf_{n\to\infty} \mathbb{P}\left(\max_{k>K_0} \hat{\theta}_L^{(k)} - \hat{q}_{1-\alpha} \frac{\hat{\sigma}^{(k)}}{\sqrt{n}} \leq \theta_L \right) = 1.$$

To do this, note that whenever $\alpha \leq 0.5$, $\hat{q}_{1-\alpha}\hat{\sigma}^{(k)} \geq 0$. Thus, it suffices to show that $\theta_L^{(k)} \leq \theta_L$ with probability 1 asymptotically. Yet Condition 2 guarantees that (i) $\hat{\beta}^{(k)} \leq \beta_L$ with probability one asymptotically and (ii) $\hat{\beta}^{(k)} - \beta_L = \omega_p(n^{-1/2})$ (see the proof of Proposition 3.2 in Appendix D.6) for $k > K_0$. Furthermore, we assume that the partial derivative of $\partial_b h(b, \kappa)$ is uniformly bounded above some constant; since h is continuously differentiable, this means that $\partial_b h(b, x)$ is uniformly bounded above some constant γ for all x in a neighborhood of κ. Since $\hat{\kappa} \xrightarrow{\text{P}} \kappa$ asymptotically, we have that with probability one asymptotically,

$$\begin{aligned} \hat{\theta}_L^{(k)} - \theta_L &= h(\hat{\beta}^{(k)}, \hat{\kappa}) - h(\beta_L, \kappa) \\ &\leq \underbrace{h(\beta_L, \hat{\kappa}) - h(\beta_L, \kappa)}_{O_p(n^{-1/2})} + \underbrace{\gamma(\hat{\beta}^{(k)} - \beta_L)}_{\text{nonpositive and } \omega_p(n^{-1/2})}. \end{aligned}$$

In particular, the left term is $O_p(n^{-1/2})$ (or smaller) by the delta method, and the right term is $\omega_p(n^{-1/2})$ by the previous remarks. Since the right-hand term dominates the left-term and is asymptotically less than zero, this implies that $\hat{\theta}_L^{(k)} - \theta_L \leq 0$ with probability one for all $k > K_0$. This completes the proof. □

B Deep Dual Bounds: an alternative approach for computation

B.1 Core Methodology

In Section 2.2, we presented our core methodology, which involves fitting conditional distributions $\hat{P}_{Y(0)|X}$ and $\hat{P}_{Y(1)|X}$, and solving the dual problem equation (11) for each $x \in \mathcal{X}$. However, this two-step approach becomes less effective when modeling conditional distributions is challenging. For instance, with complex data sources like images or texts, simple methods such as linear models or quantile regressions may lead to severe model misspecification when fitting conditional distributions. While our dual bounds remain valid, large errors in distribution fitting can significantly reduce their tightness, as indicated by Theorem 3.2.

Inspired by the recent success of deep learning in dealing with complex data and its application in optimal transport (Makkuva et al., 2020), here we present Deep Dual Bounds: an alternative approach that fits the dual bounds via end-to-end training with deep neural networks. Instead of decomposing the problem into fitting the conditional distribution and solving dual variable $\hat{\nu}_0(y, x)$ and $\hat{\nu}_1(y, x)$ for each fixed $x \in \mathcal{X}$, we aim at directly learn the optimal dual $\hat{\nu}_0(y, x)$ and $\hat{\nu}_1(y, x)$ variables on the product space of $\mathcal{Y} \times \mathcal{X}$. From the strong duality result (Theorem 2.1), we know that the partial identification bounds can be obtained via the dual formulation on the product space

$$\theta_L = \sup_{\nu_0, \nu_1 \in \mathcal{V}} \mathbb{E}_{P^\star}[\nu_0(Y(0), X) + \nu_1(Y(1), X)]. \tag{60}$$

Directly solving this problem is generally hard, since it involves constraints on potentially high dimensional X. However, weak duality (Theorem 2.1) allows us to obtain valid partial identification bounds as long as $(\hat{\nu}_0, \hat{\nu}_1) \in \mathcal{V}$. This constraint can be enforced using the strategy described in Section 4.1, allowing us to employ heuristic methods to overcome computational difficulties.

First, we can transform the constrained optimization problem in equation (60) into an unconstrained optimization problem by adding a proper penalty on the objective function. Similar to the conditional case as described in equation (30), we consider using the maximum deviation from the constraints as the penalty

in an objective. However, since the constraint contains $f(y_0, y_1, x)$, which requires us to acquire the counterfactual outcome when evaluated on data, it is not straightforward to evaluate the penalty on the data as in equation (30).

To resolve this issue without any distributional fitting, we use the matching method (Abadie and Imbens, 2006; Stuart, 2010) to impute the counterfactual outcome. For each unit (X_i, W_i, Y_i) in the treatment group (i.e., $W_i = 1$), we match it with another unit j from the control group via finding its nearest neighbor

$$j = \arg\min_{k:W_k=0} \|X_i - X_k\|,$$

and impute the $Y_i(0)$ with the observed outcome Y_j of unit j, hence we obtain a triplet sample

$$(\tilde{X}_i, \tilde{Y}_i(0), \tilde{Y}_i(1))) = (X_i, Y_i, Y_j). \tag{61}$$

We apply a similar process to impute $Y(1)$ for units in the control group. Although the matching method will suffer from distribution shifts compared to the ground-truth joint distribution P^*, it can be regarded as a good proxy when the sample size is large. In the following, we denote $\hat{P}_{\text{match}}$ as the empirical distribution of samples $\{(\tilde{X}_i, \tilde{Y}_i(0), \tilde{Y}_i(1))\}_{i=1}^n$ created by equation (61).

With the samples created by the matching method, we can use the following objective function to optimize over the whole space of $(X, Y(0), Y(1))$:

$$\begin{aligned}
&\hat{O}(\nu_0(y_0, x), \nu_1(y_1, x), \{\lambda_\ell(x)\}_{\ell=1}^L) \\
:=&\mathbb{E}_{\hat{P}_{\text{match}}}\left[\nu_0(Y(0), X) + \nu_1(Y(1), X)\right] \\
&- \max_{i=1}^n \left[\nu_0(\tilde{Y}_i(0), \tilde{X}_i) + \nu_1(\tilde{Y}_i(1), \tilde{X}_i) - \sum_{\ell=1}^L \lambda_\ell(X_i) w_{X_i,\ell}(\tilde{Y}_i(0), \tilde{Y}_i(1)) - f(\tilde{Y}_i(0), \tilde{Y}_i(1)), \tilde{X}_i)\right].
\end{aligned} \tag{62}$$

To capture the complicated structure of covariates, we can use two deep neural networks to model the dual variables and optimize them via gradient descent with respect to the objective function defined above. Regarding the constraints multiplier $\lambda_\ell(x)$, we can also use neural networks when the constraints are complicated. But in most of the applications the constraints $w_{x,\ell}$ are the same for all $x \in \mathcal{X}$ (e.g. the unconditional monotonicity constraint in Lee (2009)), or the same within some partition of $\mathcal{X}$ (e.g., the conditional monotonicity constraint in Semenova (2021)). Thus, it suffices to use constant or simple functions and optimize the parameters with gradient descent. As long as we obtain $\hat{\nu}_0(y_0, x), \hat{\nu}_1(y_1, x), \{\hat{\lambda}_\ell(x)\}_{\ell=1}^L$, we can use the same procedure as described in Section 4.1 to enforce the constraint and obtain valid partial identification bounds.

Remark 13. The purpose of the matching mechanism is to ensure we can assess the maximum deviation from the constraint as a joint function of $(X, Y(0), Y(1))$ with a lower computational cost. On a high level, it can be regarded as an analog of fitting the conditional distributions in the two-stage methods, where we make the constraints identifiable by decomposing the joint distributions into marginal distribution and conditional distribution. While we use maximum deviation on the matching sample as an intuitive penalty function, other options, such as entropy regularization, could potentially improve accuracy. Exploring better penalty functions remains a direction for future research.

Compared to two-stage methods, Deep Dual Bounds condenses the computational pipeline into a single unconstrained optimization problem, enabling end-to-end training with more powerful models like neural networks. Although the two-stage method leverages more of the conditional structure in the problem, it either suffers from model misspecification errors with simple models (e.g., linear models) or high sampling costs with advanced models (e.g., deep generative models). Therefore, deep dual methods should outperform when the underlying structure is complex.

However, it's important to note that the current deep dual objective function 62 can be challenging to optimize in practice. This is primarily due to the lack of smoothness introduced by the penalty term, which may be exacerbated when the objective function or constraints are not smooth. In contrast, the two-stage method reduces the problem to a low-dimensional space and can be efficiently optimized through linear programming with discretization. Consequently, the two-stage method remains preferable for data with relatively simple structures.

B.2 Experimental Results

Now we test the performance of the deep dual methods on the real data applications. In the following experiments, we use 5-layer fully connected ReLU neural nets to learn the dual function. We use a full batch Adam optimizer with a learning rate of 0.05, weight decay 1e-4, to optimize the deep dual objective equation (62) for 400 epochs in total. Since neural nets generalize poorly out of the data distribution, we will randomly sample 20 points from control and treatment groups to form pairs (y_0, y_1) and enforce the constraints on the samples. Similar to the experiments in Section 5, we use cross-fitting with 10 folds. The experimental results are shown in Table 4.

Dataset	Deep Dual LB	Deep Dual UB	Two Stage LB	Two Stage UB
Persuasion Effect (Section 5.1)	0.0 (0.000)	0.6416 (0.097)	0.038 (0.027)	0.365 (0.019)
401k Eligibility (Section 5.3)	12626 (1609)	69597 (6945)	5564 (1201)	47286 (1258)

Table 4: This table compares the estimated dual bounds with deep dual and two-stage methods. Details of experimental settings are described in Section 5. For two-stage methods, we only listed the model with the most sharp bounds from Table 1 and 3. Standard errors are shown in parentheses.

In Table 4, we omitted the performance on intensive margin estimation because the estimands lack smoothness and deep dual methods fail to obtain meaningful results. On the other two applications, the Deep Dual method gives comparable performance to the two-stage methods. We want to remark that the relationships between covariates and outcomes in these applications are relatively simple, therefore two-stage methods have small model misspecification errors on distributional regression and yield better performance. Moreover, we observe numerical instability regards the choice of hyperparameters due to the non-convexity of the deep dual objective. However, Deep Dual methods have the potential to deal with datasets with complicated structures, where modeling the conditional distribution can be extremely hard. We leave further investigation on the power of deep dual as an important future direction.

C Proofs from Section 2 and 4

C.1 Proof of Theorem 2.1

In this section, we prove Theorem 2.1 in two steps. The first step is to prove strong duality for the constrained optimal transport formulation in the absence of covariates. The second step is to show the problem separates in X and the Kantorovich duals can be constructed by conditioning on X. Finally, we present primitive conditions that justify the measurability of conditional Kantorovich duals with respect to X so that $\theta_L(X)$ is a random variable.

C.1.1 Step I: strong duality without covariates

The standard Monge-Kantorovich optimal transport problem can be formulated as:

$$\theta_L = \inf_P \mathbb{E}_P[f(Y(0), Y(1))] \text{ s.t. } P_{Y(1)} = P^\star_{Y(1)} \text{ and } P_{Y(0)} = P^\star_{Y(0)}.$$

We state a version of Kantorovich strong duality below for completeness. The proof can be found in Villani et al. (2009); Zaev (2015).

Definition 3. Let $Z_0, Z_1, Z = Z_0 \times Z_1$ be Polish spaces, $P^\star_{Y(0)}, P^\star_{Y(1)}$ be two probability measures on Z_0 and Z_1, define the functional spaces

$$C_L(P^\star_{Y(i)}) = \{f \in L^1(Z_i, P^\star_{Y(i)}) \cap C(Z_i)\} \text{ for } i \in \{0, 1\}$$

as the continuous and absolutely integrable functions with respect to the topology induced by the $L^1(Z_i, P^\star_{Y(i)})$ norm. For the joint space, define

$$C_L(P^\star) = \{h \in C(Z) : \exists f_0 \in C_L(P^\star_{Y(0)}), f_1 \in C_L(P^\star_{Y(1)}) \text{ s.t. } |h| \leq f_1 + f_2\}$$

Theorem C.1. *Let $Z_0, Z_1, Z = Z_0 \times Z_1$ be Polish spaces, $P^\star_{Y(0)}, P^\star_{Y(1)}$ be two probability measures on Z_0 and Z_1, $f \in C_L(P^\star)$. Define the feasible set as*

$$\mathcal{Q} := \left\{ P \in \mathcal{Q}_0 : P_{Y(1)} = P^\star_{Y(1)} \text{ and } P_{Y(0)} = P^\star_{Y(0)} \right\} \tag{63}$$

where $\mathcal{Q}_0$ denotes the set of all probability measures on Z. Then strong duality holds, that is,

$$\inf_{P \in \mathcal{Q}} \mathbb{E}_P[f(Y(0), Y(1))] = \sup_{\nu_0 + \nu_1 \le f} \mathbb{E}_{P^\star_{Y(0)}}[\nu_0(Y(0))] + \mathbb{E}_{P^\star_{Y(1)}}[\nu_1(Y(1))] \tag{64}$$

where $\nu_0 \in C_L(P^\star_{Y(0)}), \nu_1 \in C_L(P^\star_{Y(1)})$.

Remark 14. The assumption of continuity could be weakened to lower semi-continuity; see Section 5 of Villani et al. (2009) for details. Here we state the stronger version because it is needed for our next theorem on the constrained optimal transport problems.

With extra constraints, we can similarly derive the following duality theorem:

Theorem C.2. *Let $Z_0, Z_1, Z = Z_0 \times Z_1$ be Polish spaces, $P^\star_{Y(0)}, P^\star_{Y(1)}$ be two probability measures on Z_0 and Z_1, $f \in C_L(P^\star)$, and let W be a convex cone contained in $C_L(P^\star)$. Define the feasible set*

$$\mathcal{Q}_W := \left\{ P \in \mathcal{Q}_0 : P_{Y(1)} = P^\star_{Y(1)} \text{ and } P_{Y(0)} = P^\star_{Y(0)}, \mathbb{E}_P[w(Y(0), Y(1))] \le 0, \forall w \in W \right\} \tag{65}$$

where $\mathcal{Q}_0$ denotes the set of all probability measures on Z. Assume that $\mathcal{Q}_W$ is not empty, then the minimum of $\inf_{P \in \mathcal{Q}_W} \mathbb{E}_P[f(Y(0), Y(1))]$ can be achieved and strong duality holds in the sense that

$$\inf_{P \in \mathcal{Q}_W} \mathbb{E}_P[f(Y(0), Y(1))] = \sup_{w \in W} \sup_{\nu_0 + \nu_1 - w \le f} \mathbb{E}_{P^\star_{Y(0)}}[\nu_0(Y(0))] + \mathbb{E}_{P^\star_{Y(1)}}[\nu_1(Y(1))] \tag{66}$$

where $\nu_0 \in C_L(P^\star_{Y(0)}), \nu_1 \in C_L(P^\star_{Y(1)})$.

To prove Theorem C.2, we need a general version of the minimax theorem.

Theorem C.3 (Theorem 2.4.1 in Adams and Hedberg (1999))**.** *Let K be a compact convex subset of a Hausdorff topological vector space, Y be a convex subset of an arbitrary vector space, and h be a real-valued function ($\le +\infty$) on $K \times Y$, which is lower semicontinuous in x for each fixed y, convex in $x \in K$, and concave in $y \in Y$. Then*

$$\min_{x \in K} \sup_{y \in Y} h(x, y) = \sup_{y \in Y} \min_{x \in K} h(x, y)$$

With Theorem C.1 and C.3, we can prove Theorem C.2.

Proof of Theorem C.2. First, it's straightforward to prove the LHS is at least as large as the RHS:

$$\begin{aligned}
\inf_{P \in \mathcal{Q}_W} \mathbb{E}_P[f(Y(0), Y(1))] &\ge \inf_{P \in \mathcal{Q}_W} \sup_{\nu_0 + \nu_1 - w \le f} \mathbb{E}_P[\nu(Y(0)) + \nu(Y(1)) - w(Y(0), Y(1))] \\
&\ge \inf_{P \in \mathcal{Q}_W} \sup_{\nu_0 + \nu_1 - w \le f} \mathbb{E}_{P^\star_{Y(0)}}[\nu_0(Y(0))] + \mathbb{E}_{P^\star_{Y(1)}}[\nu_1(Y(1))] \\
&= \sup_{\nu_0 + \nu_1 - w \le f} \mathbb{E}_{P^\star_{Y(0)}}[\nu_0(Y(0))] + \mathbb{E}_{P^\star_{Y(1)}}[\nu_1(Y(1))].
\end{aligned}$$

To prove the other direction, we note that

$$\begin{aligned}
&\sup_{\nu_0 + \nu_1 - w \le f} \mathbb{E}_{P^\star_{Y(0)}}[\nu_0(Y(0))] + \mathbb{E}_{P^\star_{Y(1)}}[\nu_1(Y(1))] \\
&= \sup_{w \in W} \sup_{\nu_0 + \nu_1 \le f + w} \mathbb{E}_{P^\star_{Y(0)}}[\nu_0(Y(0))] + \mathbb{E}_{P^\star_{Y(1)}}[\nu_1(Y(1))] \\
&= \sup_{w \in W} \inf_{P \in \mathcal{Q}} \mathbb{E}_P[f(Y(0), Y(1)) + w(Y(0), Y(1))]
\end{aligned} \tag{67}$$

where the last equality is obtained by applying Theorem C.1 with f replaced by $f + w$. Now we can apply the minimax theorem to interchange the supermum and infimum. Specifically, let $K = \mathcal{Q}, Y = W, h(P, w) = \mathbb{E}_P[f(Y(0), Y(1)) + w(Y(0), Y(1))]$ in Theorem C.3. It is a well-known consequence of the Prokhorov theorem that the set $\mathcal{Q}$ is compact under the topology of weak convergence, and $\mathcal{Q}_W$ is a closed set of $\mathcal{Q}$, thus is

also compact. The functional h is linear in both arguments and thus it is convex in $x \in K$ and concave in $y \in Y$. Furthermore, h is continuous in w since w is integrable. By Corollary 1.5 of Zaev (2015), h is also continuous in P. Compactness and continuity together imply the existence of the solution. Moreover, the assumptions of Theorem C.3 are satisfied and therefore

$$\sup_{w\in W}\inf_{P\in\mathcal{Q}}\mathbb{E}_P[f(Y(0),Y(1))+w(Y(0),Y(1))]=\inf_{P\in\mathcal{Q}}\sup_{w\in W}\mathbb{E}_P[f(Y(0),Y(1))+w(Y(0),Y(1))]$$

For $P\notin\mathcal{Q}_W$, there exists $w_1\in W$ such that $\mathbb{E}_P[w_1(Y(0),Y(1))]>0$ by definition. Since W is a convex cone, we know $\alpha w_1\in W$ for any $\alpha\geq 0$. Letting $\alpha\to\infty$ we see that $\mathbb{E}_P[f(Y(0),Y(1))+\alpha w_1(Y(0),Y(1))]\to\infty$ thus $\sup_{w\in W}\mathbb{E}_P[f(Y(0),Y(1))+w(Y(0),Y(1))]=\infty$. This implies

$$\inf_{P\in\mathcal{Q}}\sup_{w\in W}\mathbb{E}_P[f(Y(0),Y(1))+w(Y(0),Y(1))]=\inf_{P\in\mathcal{Q}_W}\sup_{w\in W}\mathbb{E}_P[f(Y(0),Y(1))+w(Y(0),Y(1))].$$

Putting two pieces together, we obtain that

$$\begin{aligned}&\sup_{w\in W}\inf_{P\in\mathcal{Q}}\mathbb{E}_P[f(Y(0),Y(1))+w(Y(0),Y(1))]\\&=\inf_{P\in\mathcal{Q}_W}\sup_{w\in W}\mathbb{E}_P[f(Y(0),Y(1))+w(Y(0),Y(1))]\\&=\inf_{P\in\mathcal{Q}_W}\mathbb{E}_P[f(Y(0),Y(1))].\end{aligned}$$

where the last equality holds by the simple fact that $0\in W$ and $\mathbb{E}_P[w(Y(0),Y(1))]\leq 0$ for any $w\in W$ and $P\in Q_W$. Combining this identity with (67), the proof of the other direction is completed. □

C.1.2 Step II: separability in X

Recall that the problem with covariates is

$$\theta_L=\inf_{P\in\mathcal{P}}\mathbb{E}_P[f(Y(0),Y(1),X)]\text{ s.t. }P_{Y(1),X}=P^\star_{Y(1),X}\text{ and }P_{Y(0),X}=P^\star_{Y(0),X},\tag{68}$$

where

$$\mathcal{P}=\left\{\text{joint distributions }P\text{ over }\mathcal{Y}^2\times\mathcal{X}\text{ s.t. }\mathbb{E}_P[w(Y(0),Y(1))\mid X=x]\leq 0\ \ \forall w\in\mathcal{W}_x,x\in\mathcal{X}\right\},$$

and $\mathcal{W}_x=\{w_{x,1},\cdots,w_{x,L}\}$ is a finite collection of functions. In particular, by the linearity of expectation, we know that $\mathbb{E}_P[w(Y(0),Y(1))\mid X=x]\leq 0\ \forall w\in\mathcal{W}_x$, is equivalent to $\mathbb{E}_P[w(Y(0),Y(1))\mid X=x]\leq 0\ \forall w\in W_x$, where W_x is the convex cone spanned by $\{w_{x,1},\cdots,w_{x,L}\}$. Now we are ready to prove a strong duality result for equation (68).

Theorem C.4. *For θ_L as defined in equation (68), and for fixed x, we define $\theta_L(x)$ as*

$$\begin{aligned}\theta_L(x)=&\inf_P\mathbb{E}_P[f(Y(0),Y(1),x)]\\&\text{s.t. }P_{Y(0)|X=x}=P^\star_{Y(0)|X=x},P_{Y(1)|X=x}=P^\star_{Y(1)|X=x},\\&\quad\mathbb{E}_P[w(Y(0),Y(1))|X=x]\leq 0,\forall w\in W_x,\end{aligned}\tag{69}$$

Assume that for each x, $f(Y(0),Y(1),x)\in C_L(P^\star),W_x\subset C_L(P^\star)$, and there exists an optimal solution $P^{opt}_{Y(0),Y(1)|X=x}$ of the problem (69) that gives a regular conditional probability distribution. Then we have $\theta_L=\mathbb{E}_{P^\star_X}[\theta_L(X)]$.

Moreover, let

$$\nu^\star_{0,x},\nu^\star_{1,x}\in\underset{\nu_{0,x},\nu_{1,x}\in\mathcal{V}_x}{\arg\max}\ \mathbb{E}_{P^\star_{Y(0)|X=x}}[\nu_{0,x}(Y(0))]+\mathbb{E}_{P^\star_{Y(1)|X=x}}[\nu_{1,x}(Y(1))],\tag{70}$$

then $\theta_L=\mathbb{E}_{P^\star}[\nu^\star_{0,X}(Y(0))+\nu^\star_{1,X}(Y(1))]$, if $\nu^\star_{0,X}(Y(0)),\nu^\star_{1,X}(Y(1))$ are measurable with respect to $(X,Y(0),Y(1))$ and integrable under $P^\star$.

Proof. We denote by $\mathcal{Q}_x$ the set of conditional distributions that correspond to a feasible solution to problem (69). Note that

$$\begin{aligned}\theta_L&=\inf_{P\in\mathcal{P}}\mathbb{E}_P[f(Y(0),Y(1),X)]&&\text{s.t. }P_{Y(k),X}=P^\star_{Y(k),X}\text{ for }k\in\{0,1\}\\&=\inf_{P\in\mathcal{P}}\mathbb{E}_{P_X}\mathbb{E}_{P_{Y(0),Y(1)|X}}[f(Y(0),Y(1),X)]&&\text{s.t. }P_{Y(k),X}=P^\star_{Y(k),X}\text{ for }k\in\{0,1\}.\end{aligned}$$

By the constraint $P_{Y(1),X} = P^\star_{Y(1),X}$ and $P_{Y(0),X} = P^\star_{Y(0),X}$, we know that for each $P \in \mathcal{P}$, we will have $P_X = P^\star_X$ and $P_{Y(0),Y(1)|X=x} \in \mathcal{Q}_x$. Under the assumptions, there exists a regular conditional probability distribution $P^{\text{opt}}_{Y(0),Y(1)|X=x} \in \mathcal{Q}_x$ that solves equation (69) for each $x \in \mathcal{X}$. As a result,

$$\theta_L(x) = \inf_{P_{Y(0),Y(1)|X=x} \in \mathcal{Q}_x} \mathbb{E}_{P_{Y(0),Y(1)|X=x}}[f(Y(0), Y(1), x)].$$

Then $\theta_L(X)$ is measurable and hence

$$\theta_L \geq \mathbb{E}_{P^\star_X}[\theta_L(X)], \tag{71}$$

To prove equality, construct the joint distribution $P^{\text{opt}}_{X,Y(0),Y(1)} = P^\star_X \times P^{\text{opt}}_{Y(0),Y(1)|X}$. Since $P^{\text{opt}}_{Y(0),Y(1)|X=x} \in \mathcal{Q}_x$ is regular, $P^{\text{opt}}_{X,Y(0),Y(1)} \in \mathcal{P}$ is a valid feasible distribution. Thus,

$$\theta_L \leq \mathbb{E}_{P^{\text{opt}}_{X,Y(0),Y(1)}}[f(Y(0), Y(1), X)] = \mathbb{E}_{P^\star_X}[\theta_L(X)].$$

Therefore, $\theta_L = \mathbb{E}_{P^\star_X}[\theta_L(X)]$.

To prove the second result, note that Theorem C.2 implies

$$\theta_L(x) = \mathbb{E}_{P^\star_{Y(0)|X=x}}[\nu^\star_{0,x}(Y(0))] + \mathbb{E}_{P^\star_{Y(1)|X=x}}[\nu^\star_{1,x}(Y(1))].$$

Since $\nu^\star_{0,X}(Y(0)), \nu^\star_{1,X}(Y(1))$ are measurable with respect to $(X, Y(0), Y(1))$ and integrable under $P^\star$, apply the Fubini's theorem

$$\theta_L = \mathbb{E}_{P^\star_X}[\theta_L(X)] = \mathbb{E}_{P^\star_X}[\mathbb{E}_{P^\star_{Y(0)|X=x}}[\nu^\star_{0,x}(Y(0))] + \mathbb{E}_{P^\star_{Y(1)|X=x}}[\nu^\star_{1,x}(Y(1))]] = \mathbb{E}_{P^\star}[\nu^\star_{0,X}(Y(0)) + \nu^\star_{1,X}(Y(1))],$$

this finishes the proof.

□

C.1.3 Proof of Theorem 2.1

The first claim about weak duality is straightforward from the fact that $(\nu_0, \nu_1) \in \mathcal{V}$ implies

$$\nu_{0,x}(y_0) + \nu_{1,x}(y_1) \leq f(y_0, y_1, x) + \sum_{\ell=1}^{L} \lambda_{x,\ell} \cdot w_{x,\ell}(y_0, y_1)$$

and as a result, for any $P \in \mathcal{P}$

$$\begin{aligned}
&g(\nu_0, \nu_1) = \mathbb{E}_{P^\star}[\nu_{0,X}(Y(0)) + \nu_{1,X}(Y(1))] = \mathbb{E}_P[\nu_{0,X}(Y(0)) + \nu_{1,X}(Y(1))] \\
\leq &\mathbb{E}_P\left[f(Y(0), Y(1), X) + \sum_{\ell=1}^{L} \lambda_{X,\ell} \cdot w_{X,\ell}(Y(0), Y(1))\right] \\
= &\mathbb{E}_P[f(Y(0), Y(1), X)] + \mathbb{E}_{P_X}\left[\sum_{\ell=1}^{L} \lambda_{X,\ell} \cdot \mathbb{E}_P[w_{X,\ell}(Y(0), Y(1))|X]\right] \\
\leq &\mathbb{E}_P[f(Y(0), Y(1), X)].
\end{aligned} \tag{72}$$

Since it holds for all $P \in \mathcal{P}$, we conclude that $g(\nu_0, \nu_1) \leq \inf_{P \in \mathcal{P}} \mathbb{E}_P[f(Y(0), Y(1), X)] = \theta_L$. The second claim about strong duality is directly implied by Theorem C.4 with W_x being the convex cone generated by $\{w_{x,1}, \ldots, w_{x,L}\}$, assuming that all assumptions therein hold.

Remark 15. Here we remark on how our examples could satisfy the regularity conditions on f and W_x.

- For Example 1, we can redefine $Y(0)$ and $Y(1)$ as $I(Y(0) < y_0)$ and $I(Y(1) < y_1)$. Then the problem becomes discrete. Clearly, the objective function $f(Y(0), Y(1)) = I(Y(0) = Y(1) = 1)$ is bounded and continuous under the discrete topology.
- For Example 2, the objective function $f(Y(0), Y(1)) = (Y(1) - Y(0))^2$ is clearly continuous under the standard Euclidean topology in $\mathbb{R}^2$. It is bounded by $2(Y(0)^2 + Y(1)^2)$ which satisfies the integrability assumption if $Y(0), Y(1)$ have finite second moments.

- For Example 3, we can equip the space for (Y, S) by the product of the Euclidean topology on $\mathbb{R}$ and the discrete topology on $\{0, 1\}$. The objective function $f((Y(0), S(0)), (Y(1), S(1))) = (Y(1) - Y(0))I(S(1) = S(0) = 1)$ is bounded by $|Y(1)| + |Y(0)|$ which satisfies the integrability assumption if $Y(0), Y(1)$ have finite first moments. Further, the constraint function $w((Y(0), S(0)), (Y(1), S(1))) = \mathbb{I}(S(0) \leq S(1))$ is continuous and integrable under the discrete topology.
- For Example 4, if the distribution of $(Y(0), Y(1))$ is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^2$, then the estimand can be equivalently formulated as $\mathbb{E}_P[\mathbb{I}(Y_i(1) - Y_i(0) < t)]$. The objective function $f(Y(0), Y(1)) = \mathbb{I}(Y(1) - Y(0) < t)$ is lower semi-continuous. Since no constraint is involved, we can apply the stronger version of Theorem C.4 discussed in Remark 14 to obtain strong duality.
- Examples 5 and 6 can be reasoned similarly as above.

C.1.4 Primitive assumptions for measurability

In Theorem C.4, we assume that the primal solution $P^{\text{opt}}_{Y(0),Y(1)|X=x}$ gives a regular conditional probability distribution, and the dual solution $\nu^\star_{0,X}(Y(0)), \nu^\star_{1,X}(Y(1))$ are measurable with respect to $(X, Y(0), Y(1))$. and integrable on the product spaces. The integrability assumption is to ensure that the bound is finite, so we skip the discussion on it. Instead, in this section, we provide primitive conditions to justify the measurability.

We remind the readers that a conditional distribution $P^{\text{opt}}_{Y(0),Y(1)|X=x}$ is a regular conditional probability distribution assumption if and only if

1. For any fixed x, $P^{\text{opt}}_{Y(0),Y(1)|X=x}(\cdot)$ is a probability distribution.
2. For any fixed $A \in \mathcal{F}$, $P^{\text{opt}}_{Y(0),Y(1)|X=x}(A)$ is a measurable function with respect to x, where $\mathcal{F}$ is the σ-algebra on the product space $\mathcal{Y}^2$.

We prove the following result that the measurability assumptions are satisfied when $\mathcal{X}$ is Euclidean and $\mathcal{Y}$ is discrete.

Proposition C.1. *Assume that $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{Y} = \{y_1, \ldots, y_K\}$, both equipped with Borel σ-algebra. Further assume that $(w_x(y_j, y_k))_{j,k\in[K]}$ is measurable in x for all $i, j \in [K]$, and for each x, the feasible set of (69) is non-empty. Then the measurability assumptions of Theorem C.4 are satisfied.*

To prove Proposition C.1, we will need the following results from the theory of linear programming (Matoušek and Gärtner, 2007):

Definition 4. A basic feasible solution of the linear program

$$\max c^T x \quad \text{s.t. } Ax = b \text{ and } x \geq \mathbf{0}$$

is a feasible solution $x \in \mathbb{R}^n$ for which there exists an m-element set $B \subseteq \{1, 2, \ldots, n\}$ such that

- the (square) matrix A_B is nonsingular, i.e., the columns indexed by B are linearly independent,
- $x_j = 0$ for all $j \notin B$.

Lemma C.5 (Theorem 4.2.3 of Matoušek and Gärtner (2007))**.** *Consider the following linear program*

$$\textit{max } c^T x \quad \textit{s.t. } Ax = b \textit{ and } x \geq \mathbf{0}.$$

1. *If there is at least one feasible solution and the objective function is bounded from above on the set of all feasible solutions, then there exists an optimal solution.*
2. *If an optimal solution exists, then there is a basic feasible solution that is optimal.*

Proof of Proposition C.1. Since $\mathcal{Y}$ is discrete and the optimization problem (69) depends on x only through the constraints $P_{Y(0)|X=x} = P^\star_{Y(0)|X=x}, P_{Y(1)|X=x} = P^\star_{Y(1)|X=x}$ and $\mathbb{E}_P[w_{x,l}(Y(0), Y(1))|X = x] \leq 0, \forall l =$

$1, \cdots, L$, we can express $P_{Y(0),Y(1)|X=x}$ as a K^2 dimensional vector, denoted by p, and write equation (69) as a linear program

$$\theta_L(x) = \min_{p \in \mathbb{R}^{K^2}} c_x^T p \quad \text{s.t. } Ap = b_x, p \geq 0, C_x p \leq 0, \tag{73}$$

where $c_x \in \mathbb{R}^{K^2}$ is the vectorization of $(f(y_j, y_k, x))_{j,k\in[K]}$, $b_x \in \mathbb{R}^{2K}$ is the concatenation of $(P(Y(0) = y_j) \mid X = x))_{j=1}^K$ and $(P(Y(1) = y_j) \mid X = x))_{j=1}^K$, and $C_x \in \mathbb{R}^{L\times K^2}$ with each row encodes the vectorization of $(w_x(y_j, y_k))_{j,k\in[K]}$. Clearly, c_x is measurable with respect to x. Since the measurable spaces of $(X, Y(0))$ and $(X, Y(1))$ are Radon spaces, b_x is measurable with respect to x. Under the assumption, C_x is also measurable with respect to x.

By introducing slack variables $s = -C_x p \in \mathbb{R}^L$, and $q = (p, s)^T \in \mathbb{R}^{K^2+L}$, we could transform it into a standard form

$$\theta_L(x) = \max_{q \in \mathbb{R}^{K^2+L}} \tilde{c}_x^T q \quad \text{s.t. } \tilde{A}_x q = \tilde{b}_x, q \geq 0, \tag{74}$$

where $\tilde{c}_x, \tilde{A}_x, \tilde{b}_x$ are measurable functions of c_x, A, b_x, C_x. Thus, they are measurable with respect to x. Moreover, by assuming the feasible set is non-empty, we can make A_x have a full row rank by removing some rows without changing the linear program. Thus, we will assume A_x has full row rank for simplicity.

For each subset $B \subset \{1, 2, \cdots, K^2 + L\}$ such that $\tilde{A}_{x,B}$ is square and nonsingular, the corresponding basic feasible solution exists if and only if $\tilde{A}_{x,B}^{-1} b_x \geq 0$. For any $b \in \mathcal{B}$ where $\mathcal{B} \subset \mathbb{R}^{2K}$ is the domain of b (namely the concatenation of two K-dimensional simplexes), we define

$$S_x(b) = \{B \subset \{1, 2, \cdots, K^2 + L\} : \tilde{A}_{x,B} \text{ is square, nonsingular and } \tilde{A}_{x,B}^{-1} b \geq 0\}.$$

Note that $S_x(b)$ can only take finitely many (set) values, denoted by $S^1, \ldots S^I$. This defines a partition $(M_{x,1}, \ldots, M_{x,I})$ of $\mathcal{B}$ where

$$M_{x,i} = \{b \in \mathcal{B} : S_x(b) = S^i\}.$$

Clearly, each $M_{x,i}$ is a polytope determined by finitely many linear inequalities whose coefficients are measurable with respect to x. Thus, $\{\mathbb{I}(\tilde{b}_x \in M_{x,i})\}_{i=1}^I$ is measurable with respect to x as well. For each $i \in \{1, \ldots, I\}$, if $\tilde{b}_x \in M_{x,i}$, then

$$\theta_L(x) = \max_{B \in S^i} \tilde{c}_x^T \tilde{A}_{x,B}^{-1} \tilde{b}_x.$$

Since it is defined over a finite number of sets, the maximum can be achieved. Denote by $B_x^{(i)}$ the maximizer (with the smallest i when multiple optimums exist). Note that $B_x^{(i)}$ is maximizer of finitely many measurable functions of $(\tilde{c}_x, \tilde{A}_x, \tilde{b}_x)$, it is also measurable with respect to x. As a result,

$$p_x^\star = \sum_{i=1}^{I_x} \tilde{A}_{x,B_x^{(i)}}^{-1} \tilde{b}_x \cdot \mathbb{I}(\tilde{b}_x \in M_{x,i}),$$

is measurable with respect to x. Thus, we have constructed a primal solution that is a regular conditional probability distribution.

Now we move to the measurability of the dual solutions. The Lagrangian dual problem of (74) is

$$\theta_L(x) = \min_{\nu} \tilde{b}_x^T \nu \quad \text{s.t. } \tilde{A}_x^T \nu \geq \tilde{c}_x. \tag{75}$$

By reparametrizing $\nu = \mu_+ - \mu_-$ for some $\mu_+ \geq 0, \mu_- \geq 0$ and setting $d = \tilde{A}_x^T \nu - \tilde{c}_x$, we can transform equation (75) into the standard form in Lemma C.5. Using the same argument for the primal solution, we can show the existence of dual variables that are measurable with respect to x.

□

Remark 16. If $\mathcal{Y}$ is continuous, a weaker result has been shown that, under some regularity assumptions, there exists a primal solution that is measurable with respect to the Borel σ-algebra generated by the weak topology (Bogachev and Malofeev, 2020; Bogachev, 2022). We expect the same technique can be used to prove the existence of a primal solution that is a regular conditional probability distribution and the measurability of dual variables, though we leave formal proof of these claims for future research.

C.2 Proof of Proposition 4.1

Now we turn to prove Proposition 4.1.

Proposition C.2. *Suppose $\hat{\nu}_{0,x}^{\text{init}}, \hat{\nu}_{1,x}^{\text{init}} : \mathcal{Y} \to \mathbb{R}$ and $\hat{\lambda}_{x,1}, \ldots, \hat{\lambda}_{x,L} \geq 0$ maximize the objective function $O(\nu_{0,x}, \nu_{1,x}, \{\lambda_{x,\ell}\}_{\ell=1}^{L})$ among all functions $\nu_{0,x}, \nu_{1,x} : \mathcal{Y} \to \mathbb{R}$ and constants $\lambda_{x,1}, \ldots, \lambda_{x,L} \geq 0$. Let c_x be the minimum constant such that $(\hat{\nu}_{0,x}^{\text{init}} - c_x, \hat{\nu}_{1,x}^{\text{init}} - c_x) \in \mathcal{V}_x$ are conditionally valid dual variables. Then $\hat{\nu}_{0,x}^{\text{init}} - c_x, \hat{\nu}_{1,x}^{\text{init}} - c_x$ solve the conditional dual problem Eq. (28).*

Proof. Denote the optimal dual variables of Eq. (28) as $\nu_0^\star, \nu_1^\star$. Note that as long as $(\hat{\nu}_{0,x}^{\text{init}} - c_x, \hat{\nu}_{1,x}^{\text{init}} - c_x) \in \mathcal{V}_x$ are conditionally valid dual variables, we always have

$$g(\hat{\nu}_{0,x}^{\text{init}} - c_x, \hat{\nu}_{1,x}^{\text{init}} - c_x)) \leq g(\nu_0^\star, \nu_1^\star)$$

where g is defined in equation (8). Suppose for the sake of contradiction that the inequality holds strictly. By the definition of c_x,

$$\max_{y_1, y_0 \in \mathcal{Y}} (\hat{\nu}_{0,x}^{\text{init}}(y_0) - c_x) + (\hat{\nu}_{1,x}^{\text{init}}(y_1) - c_x) - \sum_{\ell=1}^{L} \hat{\lambda}_{x,\ell} w_{x,\ell}(y_0, y_1) - f(y_0, y_1, x) = 0.$$

Since $\nu_{0,x}^\star, \nu_{1,x}^\star$ are both valid dual variables,

$$\max_{y_1, y_0 \in \mathcal{Y}} \nu_{0,x}^\star(y_0) + \nu_{1,x}^\star(y_1) - \sum_{\ell=1}^{L} \hat{\lambda}_{x,\ell} w_{x,\ell}(y_0, y_1) - f(y_0, y_1, x) \leq 0.$$

Furthermore, notice that subtracting a constant off from $\hat{\nu}_{0,x}^{\text{init}}, \hat{\nu}_{1,x}^{\text{init}}$ doesn't change the value of $O(\hat{\nu}_{0,x}^{\text{init}}, \hat{\nu}_{1,x}^{\text{init}}, \{\hat{\lambda}_{x,\ell}\}_{\ell=1}^{L})$. Thus if $g(\hat{\nu}_{0,x}^{\text{init}} - c_x, \hat{\nu}_{1,x}^{\text{init}} - c_x) < g(\nu_{0,x}^\star, \nu_{1,x}^\star)$, we can conclude that

$$\begin{aligned}
&O(\hat{\nu}_{0,x}^{\text{init}}, \hat{\nu}_{1,x}^{\text{init}}, \{\hat{\lambda}_{x,\ell}\}_{\ell=1}^{L}) = O(\hat{\nu}_{0,x}^{\text{init}} - c_x, \hat{\nu}_{1,x}^{\text{init}} - c_x, \{\hat{\lambda}_{x,\ell}\}_{\ell=1}^{L}) \\
=&g(\hat{\nu}_{0,x}^{\text{init}} - c_x, \hat{\nu}_{1,x}^{\text{init}} - c_x) - \max_{y_1, y_0 \in \mathcal{Y}} (\hat{\nu}_{0,x}^{\text{init}}(y_0) - c_x) + (\hat{\nu}_{1,x}^{\text{init}}(y_1) - c_x) - \sum_{\ell=1}^{L} \hat{\lambda}_{x,\ell} w_{x,\ell}(y_0, y_1) - f(y_0, y_1, x) \\
<&g(\nu_{0,x}^\star, \nu_{1,x}^\star) - \max_{y_1, y_0 \in \mathcal{Y}} \nu_{0,x}^\star(y_0) + \nu_{1,x}^\star(y_1) - \sum_{\ell=1}^{L} \hat{\lambda}_{x,\ell} w_{x,\ell}(y_0, y_1) - f(y_0, y_1, x) \\
=&O(\nu_{0,x}^\star, \nu_{1,x}^\star, \{\hat{\lambda}_{x,\ell}\}_{\ell=1}^{L})
\end{aligned}$$

which violates the definition of $(\hat{\nu}_{0,x}^{\text{init}}, \hat{\nu}_{1,x}^{\text{init}})$ as the minimizer of $O(\nu_{0,x}, \nu_{1,x}, \{\hat{\lambda}_{x,\ell}\}_{\ell=1}^{L})$. Thus we must have equality

$$g(\hat{\nu}_{0,x}^{\text{init}} - c_x, \hat{\nu}_{1,x}^{\text{init}} - c_x)) = g(\nu_0^\star, \nu_1^\star).$$

□

D Proofs from Section 3

D.1 Main proofs from Section 3.1

Although we give a proof sketch of Theorems 3.1 and 3.5 in Section 3.1, we give a few more details here for the sake of completeness. We also prove Corollary 3.1.

Theorem 3.1. *Assume Assumption 3.1. For any $B \geq 0$, let $\mathcal{P}_B \subset \mathcal{P}$ denote the set of all laws $P \in \mathcal{P}$ such that $\hat{\nu}$ satisfies Assumption 3.2 under P. Then*

$$\liminf_{n \to \infty} \inf_{P \in \mathcal{P}_B} \mathbb{P}(\hat{\theta}_{\text{LCB}} \leq \theta_L) \geq 1 - \alpha.$$

Proof. As in Section 3.1, we begin with the decomposition

$$\theta_L - \hat{\theta}_{\text{LCB}} = \underbrace{\theta_L - \tilde{\theta}_L}_{\text{Term A}} + \underbrace{\tilde{\theta}_L - \hat{\theta}_{\text{LCB}}}_{\text{Term B}}.$$

Term A is positive deterministically by weak duality. To analyze Term B, let $S_i = \frac{\hat{\nu}_1(Y_i,X_i)W_i}{\pi(X_i)} + \frac{\hat{\nu}_0(Y_i,X_i)(1-W_i)}{1-\pi(X_i)}$ for $i \in \mathcal{D}_2$ and let $\hat{\sigma}_s$ be the sample standard deviation of $\{S_i\}_{i\in\mathcal{D}_2}$. Now, by construction,

$$\begin{aligned}\tilde{\theta}_L - \hat{\theta}_{\text{LCB}} &= \mathbb{E}\left[\hat{\nu}_0(Y_i(0), X_i) + \hat{\nu}_1(Y_i(1), X_i) \mid \mathcal{D}_1\right] - \frac{1}{n_2}\sum_{i=1}^{n} S_i + \Phi^{-1}(1-\alpha)\frac{\hat{\sigma}_s}{\sqrt{n_2}} \\ &= \mathbb{E}[S_i \mid \mathcal{D}_1] - \frac{1}{n_2}\sum_{i=1}^{n} S_i + \Phi^{-1}(1-\alpha)\frac{\hat{\sigma}_s}{\sqrt{n_2}}.\end{aligned}$$

One approach to analyze this sum would be to apply the standard univariate CLT conditional on $\mathcal{D}_1$ and to let n_2 grow to ∞. However, we must be slightly careful, because the rate of the convergence of the CLT depends on (e.g.) the higher moments of S_i, which depend on $\hat{\nu}$, and $\hat{\nu}$ changes with n (since $\mathcal{D}_1$ changes with n). Instead, we will apply the Lyapunov CLT for triangular arrays.

Indeed, Assumption 3.2 specifies that the conditional variance of S_i is bounded away from zero and its fourth conditional moment is uniformly bounded. (In particular, the latter follows because the fourth conditional moment of $\hat{\nu}(Y(k), X)$ is uniformly bounded and because of strict overlap, i.e., that $\pi(X) \in [\Gamma, 1-\Gamma]$ for some $\Gamma > 0$.) This moment condition, in combination with the fact that $\{S_i\}_{i\in\mathcal{D}_2}$ are i.i.d. conditional on $\mathcal{D}_1$, allows us to apply the Lyapunov CLT conditionally on $\mathcal{D}_1$:

$$\frac{\sqrt{n_2}}{\text{Var}(S_i \mid \mathcal{D}_1)}\left(\frac{1}{n_2}\sum_{i=1}^{n}(S_i - \mathbb{E}[S_i \mid \mathcal{D}_1])\right) \xrightarrow{\text{d}} \mathcal{N}(0,1).$$

Note that the Lyapunov CLT holds for any triangular array of random variables as long as the moment condition from Assumption 3.2 holds; therefore, this convergence is uniform over $\mathcal{P}_B$. A similar argument based on the law of large numbers for triangular arrays implies that $\hat{\sigma}_s \xrightarrow{\text{p}} \sqrt{\text{Var}(S_i \mid \mathcal{D}_1)}$ as $n \to \infty$, and furthermore that this convergence is uniform over $\mathcal{P}_B$.[10] Then, Slutsky's theorem implies that

$$\frac{\sqrt{n_2}}{\hat{\sigma}_s}\left(\frac{1}{n_2}\sum_{i=1}^{n}(S_i - \mathbb{E}[S_i \mid \mathcal{D}_1])\right) \xrightarrow{\text{d}} \mathcal{N}(0,1).$$

This proves that $\liminf_{n\to\infty} \inf_{P\in\mathcal{P}_B} \mathbb{P}(\tilde{\theta}_L - \hat{\theta}_{\text{LCB}} \geq 0) = 1-\alpha$, completing the proof. □

Remark 17. We can substantially relax Assumption 3.2 without changing the proof of Theorem 3.1. In fact, all we need to apply the Lyapunov CLT is that

$$\frac{\mathbb{E}_P[|S_i|^{2+\delta} \mid \mathcal{D}_1]}{\text{Var}_P(S_i \mid \mathcal{D}_1)} \leq Bn^{\delta/2-\epsilon} \tag{76}$$

holds for some $\epsilon > 0, \delta > 0, B > 0$. Furthermore, this does not need to hold with probability one: instead, we could require that

$$\mathbb{P}_P\left(\frac{\mathbb{E}_P[|S_i|^{2+\delta} \mid \mathcal{D}_1]}{\text{Var}_P(S_i \mid \mathcal{D}_1)} \leq Bn^{\delta/2-\epsilon}\right) \geq 1 - a_n \text{ for all } n \in \mathbb{N} \tag{77}$$

for some deterministic sequence $a_n \to 0$. Then, asymptotically, we can apply the Lyapunov CLT conditional on $\mathcal{D}_1$ with probability uniformly approaching one for any distribution such that the previous equation is satisfied.

The only other time we use Assumption 3.2 is to show $\hat{\sigma}_s \xrightarrow{\text{p}} \sqrt{\text{Var}(S_i \mid \mathcal{D}_1)}$ holds uniformly over $\mathcal{P}_B$. Here, we can again replace Assumption 3.2 with any assumption guaranteeing uniform convergence (in probability) of $\hat{\sigma}_s$.

Next, we prove Corollary 3.1. To do this, it is helpful to review a result from Chernozhukov et al. (2018b) (labeled as Theorem 4.3 in the original paper).

Proposition D.1 (Chernozhukov et al. (2018b))**.** *Let $X_1, \ldots, X_n \in \mathbb{R}^p$ be i.i.d. variables satisfying the following:*

1. *$\mu \coloneqq \mathbb{E}[X_1] \leq 0$ holds elementwise.*

[10] In particular, since the fourth moment of S_i given $\mathcal{D}_1$ is bounded, Chebyshev's inequality implies the uniform convergence of $\hat{\sigma}_s^2 \xrightarrow{\text{p}} \text{Var}(S_i \mid \mathcal{D}_1)$.

2. Let $Z_i \coloneqq X_i - \mu$ be the centered variables and define

$$B_n \coloneqq \max_{j\in[p]} \max(\mathbb{E}[|Z_{1j}|^4]^{1/2}, \mathbb{E}[|Z_{ij}|^3]) + \mathbb{E}[\max_{j\in[p]} |Z_{1j}|^4]^{1/4} < \infty.$$

Let $\hat{\mu}_j = \frac{1}{n}\sum_{i=1}^n X_{ij}$ and $\hat{\sigma}_j = \frac{1}{n}\sum_{i=1}^n (X_{ij} - \hat{\mu}_j)^2$. Also, define the statistic

$$T = \max_{j\in[p]} \frac{\sqrt{n}\hat{\mu}_j}{\hat{\sigma}_j}$$

and let

$$T^{(b)} = \max_{j\in[p]} \frac{n^{-1/2}\sum_{i\in[n]} W_i(X_i - \hat{\mu}_j)}{\hat{\sigma}_j}$$

for $W_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$. Let $\hat{q}_{1-\alpha}$ denote the conditional $1-\alpha$ quantile of $T^{(b)}$ given $X_1, \dots, X_n$. Then if there exist constants $c_1 \in (0, 1/2), C_1 \geq 0$ such that

$$B_n^2 \log(pn)^{7/2} \leq C_1 n^{1/2-c_1}, \tag{78}$$

we have that there exist constants $c, C > 0$ depending only on c_1, C_1 such that

$$\mathbb{P}(T \geq \hat{q}_{1-\alpha}) \leq \alpha + Cn^{-c} \to \alpha.$$

Assumption 3.3. *For K estimates $\hat{\nu}^{(1)}, \dots, \hat{\nu}^{(K)}$ of $\nu^\star$, for $i \in \mathcal{D}_2$, define the IPW summands*

$$S_i^{(k)} \coloneqq \frac{\hat{\nu}_1^{(k)}(Y_i, X_i)W_i}{\pi(X_i)} + \frac{\hat{\nu}_0^{(k)}(Y_i, X_i)(1-W_i)}{1-\pi(X_i)} \text{ and } Z_{ik} = S_i^{(k)} - \mathbb{E}[S_i^{(k)} \mid \mathcal{D}_1].$$

We assume there exists $\epsilon \in (0, 1/4), c > 0$ such that

$$B_n \coloneqq \max_{k\in[K]} \left((\mathbb{E}[|Z_{ik}|^4 \mid \mathcal{D}_1)^{1/2} \vee (\mathbb{E}|Z_{ik}|^3 \mid \mathcal{D}_1)\right) + \mathbb{E}\left[\max_{k\in[K]} |Z_{ik}|^4 \mid \mathcal{D}_1\right]^{1/4} \leq c\frac{n^{1/4-\epsilon}}{\log(Kn)^{7/4}}.$$

Corollary 3.1. *Suppose the analyst computes K estimates $\hat{\nu}^{(1)}, \dots, \hat{\nu}^{(K)}$ of $\nu^\star$ on $\mathcal{D}_1$ and uses the multiplier bootstrap to compute a lower bound $\hat{\theta}_{\text{LCB}}^{\text{MB}}$ as defined in Def. 2. Fix $c > 0, \epsilon \in (0, 1/4)$ and let $\mathcal{P}_{c,\epsilon}$ denote the set of laws $P \in \mathcal{P}$ such that Assumption 3.3 holds. Then under Assumption 3.1,*

$$\liminf_{n\to\infty} \inf_{P\in\mathcal{P}_{c,\epsilon}} \mathbb{P}(\hat{\theta}_{\text{LCB}}^{\text{MB}} \leq \theta_L) \geq 1 - \alpha.$$

Proof. The proof is a straightforward application of Proposition D.1 and weak duality, but for completeness, we will state it here. Recall the notation from Section 2.3: we let $\tilde{\theta}_L^{(k)} \coloneqq g(\hat{\nu}^{(k)})$, define $\tilde{\theta}_L = \max_{k\in[K]} \tilde{\theta}_L^{(k)}$, and define the summands

$$S_i^{(k)} \coloneqq \frac{\hat{\nu}_1^{(k)}(Y_i, X_i)W_i}{\pi(X_i)} + \frac{\hat{\nu}_0^{(k)}(Y_i, X_i)(1-W_i)}{1-\pi(X_i)} \text{ for } k \in [K].$$

We also set $\bar{S}_k$ and $\hat{\sigma}_k^2$ to be the empirical mean and variance of $\{S_i^{(k)} : i \in \mathcal{D}_2\}$:

$$\bar{S}_k = \frac{1}{\mathcal{D}_2}\sum_{i\in\mathcal{D}_2} S_i^{(k)} \text{ and } \hat{\sigma}_k^2 = \frac{1}{|\mathcal{D}_2|}\sum_{i\in\mathcal{D}_2} (S_i^{(k)} - \bar{S}_k)^2.$$

For any $a \in \mathbb{R}$, define the test statistic

$$T(a) \coloneqq \max_{k\in[K]} \frac{\sqrt{|\mathcal{D}_2|}(\bar{S}_k - a)}{\hat{\sigma}_k}.$$

We also define the *multiplier bootstrap variant*: for $W_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$, we define

$$T^{(b)}(a) \coloneqq \max_{k\in[K]} \frac{|\mathcal{D}|_2^{-1/2}\sum_{i\in\mathcal{D}_2} W_i(S_i^{(k)} - a - (\bar{S}_k - a))}{\hat{\sigma}_k} = T^{(b)}$$

where we note that $T^{(b)}(a)$ does not depend on a, so we abbreviate it by $T^{(b)}$. Let $\hat{q}_{1-\alpha} \coloneqq Q_{1-\alpha}(T^{(b)} \mid \mathcal{D})$ denote the conditional quantile of $T^{(b)}$ given all the data.

Note that for any $a \geq \tilde{\theta}_L$, $\mathbb{E}[\bar{S} - a\mathbf{1}_K \mid \mathcal{D}_1] \leq 0$ holds elementwise. This, combined with Assumption 3.3, allows us to apply Proposition D.1 conditional on $\mathcal{D}_1$. Thus, there exist universal constants $c, C > 0$ such that

$$\sup_{P \in \mathcal{P}_{c,\epsilon}} \mathbb{P}_P \left(T(\tilde{\theta}_L) \geq \hat{q}_{1-\alpha} \mid \mathcal{D}_1 \right) \leq \alpha + Cn^{-c}. \tag{79}$$

Applying the tower property and taking limits yields

$$\limsup_{n \to \infty} \sup_{P \in \mathcal{P}_{c,\epsilon}} \mathbb{P}_P(T(\tilde{\theta}_L) \geq \hat{q}_{1-\alpha}) \leq \alpha. \tag{80}$$

Recall also the definition of $\hat{\theta}^{\mathrm{MB}}_{\mathrm{LCB}}$ from Eq. (16):

$$\begin{aligned}
\hat{\theta}^{\mathrm{MB}}_{\mathrm{LCB}} &\coloneqq \max_{k \in [K]} \bar{S}_k - \hat{q}_{1-\alpha} \frac{\hat{\sigma}_k}{\sqrt{|\mathcal{D}_2|}} \\
&= \max \left\{ a \in \mathbb{R} : \max_{k \in [K]} \frac{\sqrt{|\mathcal{D}_2|}(\bar{S}_k - a)}{\hat{\sigma}_k} \geq \hat{q}_{1-\alpha} \right\} \\
&= \max \left\{ a \in \mathbb{R} : T(a) \geq Q_{1-\alpha}(T^{(b)} \mid \mathcal{D}) \right\}.
\end{aligned}$$

Since $T(a)$ is decreasing in a, we have that

$$\hat{\theta}^{\mathrm{MB}}_{\mathrm{LCB}} \geq \tilde{\theta}_L \Leftrightarrow T(\tilde{\theta}_L) \geq Q_{1-\alpha}(T^{(b)} \mid \mathcal{D}).$$

Therefore, by Eq. (80), we conclude

$$\limsup_{n \to \infty} \sup_{P \in \mathcal{P}_{c,\epsilon}} \mathbb{P}_P(\hat{\theta}^{\mathrm{MB}}_{\mathrm{LCB}} \geq \tilde{\theta}_L) \leq \limsup_{n \to \infty} \sup_{P \in \mathcal{P}_{c,\epsilon}} \mathbb{P}_P(T(\tilde{\theta}_L) \geq \hat{q}_{1-\alpha}) \leq \alpha.$$

Since $\tilde{\theta}_L \leq \theta_L$ by weak duality, this completes the proof.

□

Now we prove Theorem 3.5.

Theorem 3.5. *Suppose Assumption 3.1 holds except that the propensity scores are not known. For $k \in \{0,1\}$, assume the fourth moments $\mathbb{E}[|\hat{\nu}_k(Y(k), X)|^4 \mid \mathcal{D}_1] \leq B < \infty$ and $\mathbb{E}[|\hat{c}_k(X)|^4 \mid \mathcal{D}_1] \leq B < \infty$ are uniformly bounded. Finally, assume that the estimated propensity scores $\hat{\pi}(X_i)$ are uniformly bounded away from zero and one.*

Let $\mathrm{error}_n(\hat{\pi}) \coloneqq \mathbb{E}[(\hat{\pi}(X) - \pi(X))^2 \mid \mathcal{D}_1]^{1/2}$ denote the ℓ_2 error in estimating the propensity scores and let $\mathrm{error}_n(\hat{c}) = \max_{k \in \{0,1\}} \mathbb{E}[(\hat{c}_k(X) - c_k(X))^2 \mid \mathcal{D}_1]^{1/2}$ denote the ℓ_2 error in estimating the conditional mean of $\hat{\nu}$, where X is an independent draw from the law of X_i.

Consider the two conditions below:

- *Condition 1:* $\mathrm{error}_n(\hat{\pi}) = o_{L_2}(1)$, $\mathrm{error}_n(\hat{c}) = o_{L_2}(1)$, *and the "risk-decay" condition holds:*

  $$\mathbb{E}[\mathrm{error}_n(\hat{\pi})^2]\mathbb{E}[\mathrm{error}_n(\hat{c})^2] = o(1/n).$$

  Furthermore, if $\tilde{S}_i = W_i \frac{\hat{\nu}_1(Y_i, X_i) - c_1(X_i)}{\pi(X_i)} + (1 - W_i) \frac{\hat{\nu}_0(Y_i, X_i) - c_0(X_i)}{1 - \pi(X_i)} + c_1(X_i) + c_0(X_i)$, *we assume* $\mathrm{Var}(\tilde{S}_i \mid \mathcal{D}_1) \geq \frac{1}{B}$ *is bounded away from zero.*

- *Condition 2: the outcome model is sufficiently misspecified such that the first-stage bias $\tilde{\theta}_L - \theta_L$ dominates either* $\mathrm{error}_n(\hat{\pi})$ *or* $\mathrm{error}_n(\hat{c})$. *More precisely, assume*

  $$\frac{\min(\mathrm{error}_n(\hat{c}), \mathrm{error}_n(\hat{\pi}))}{\tilde{\theta}_L - \theta_L} \xrightarrow{\mathrm{p}} 0.$$

If either Condition 1 or Condition 2 holds, then $\hat{\theta}^{\mathrm{aug}}_{\mathrm{LCB}}$ is asymptotically valid:

$$\liminf_{n \to \infty} \mathbb{P}(\hat{\theta}^{\mathrm{aug}}_{\mathrm{LCB}} \leq \theta_L) \geq 1 - \alpha.$$

Proof. This proof follows entirely from the standard theory of the AIPW estimator. We make only minor adjustments (proved in Appendix D.2) to account for the fact that $\hat{\nu}$ may change with n.

Analysis of Condition 1: Under Condition 1, standard results about the AIPW estimator (see, e.g., Wager, 2020) imply that

$$\sqrt{n}(\hat{\theta}^{\text{aug}}_{\text{LCB}} - \tilde{\theta}^{\text{aug}}_{\text{LCB}}) \xrightarrow{\text{p}} 0, \tag{81}$$

where $\tilde{\theta}^{\text{aug}}_{\text{LCB}}$ is defined equivalently to $\hat{\theta}^{\text{aug}}_{\text{LCB}}$ but with $\hat{c}_1, \hat{c}_0, \hat{\pi}$ replaced with c_1, c_0, π (respectively). See Lemma D.1 for a formal proof of this result in this setting. We note that $\liminf_n \mathbb{P}(\tilde{\theta}^{\text{aug}}_{\text{LCB}} \leq \theta_L) \geq 1-\alpha$ holds using exactly the same proof as Theorem 3.1, and since $\tilde{\theta}^{\text{aug}}_{\text{LCB}} - \theta_L = \Omega_p(n^{-1/2})$, the additional fluctuations between $\hat{\theta}^{\text{aug}}_{\text{LCB}}$ and $\hat{\theta}_{\text{LCB}}$ are asymptotically negligible.[11] Therefore, the result now follows directly from the argument in Theorem 3.1.

Analysis of Condition 2: Roughly speaking, Condition 2 tells us that $\hat{\theta}^{\text{aug}}_{\text{LCB}} \approx \tilde{\theta}_L$ and that the first-stage bias $\tilde{\theta}_L - \theta_L$ is of higher order than the fluctuations of $\hat{\theta}^{\text{aug}}_{\text{LCB}}$ around $\tilde{\theta}_L$. This follows from the standard theory of the double-robustness of the AIPW estimator (e.g. Robins et al., 1994; Wager, 2020)).

Formally, Lemma D.2 shows that

$$\hat{\theta}^{\text{aug}}_{\text{LCB}} - \tilde{\theta}_L = O_p(n^{-1/2} + \min(\text{error}_n(\hat{\pi}), \text{error}_n(\hat{c}))).$$

We note that it suffices to consider the case where $\min(\text{error}_n(\hat{\pi}), \text{error}_n(\hat{c})) = \Omega_p(n^{-1/2})$. To see this, note that if $\min(\text{error}_n(\hat{\pi}), \text{error}_n(\hat{c})) = o_p(n^{-1/2})$, then $\text{error}_n(\hat{\pi}) \cdot \text{error}_n(\hat{c}) = o_p(n^{-1/2})$ holds and Condition 1 holds, in particular because both $\text{error}_n(\hat{\pi})$ and $\text{error}_n(\hat{c})$ are uniformly bounded by the moment conditions in the theorem. In particular, $\text{error}_n(\hat{\pi}) \coloneqq \mathbb{E}[(\hat{\pi}(X) - \pi(X))^2 \mid \mathcal{D}_1]^{1/2} \leq 1$ is uniformly bounded because $\pi(X), \hat{\pi}(X) \in (0,1)$, and $\text{error}_n(\hat{c}) = \max_{k\in\{0,1\}} \mathbb{E}[(\hat{c}_k(X) - c_k(X))^2 \mid \mathcal{D}_1]^{1/2}$ is uniformly bounded because we assume $\mathbb{E}[|\hat{c}_k(X)|^{2+\delta} \mid \mathcal{D}_1]$ and $\mathbb{E}[|c_k(X)|^{2+\delta} \mid \mathcal{D}_1] \leq \mathbb{E}[|\hat{\nu}_k(Y(k), X)|^{2+\delta} \mid \mathcal{D}_1]$ are uniformly bounded for, e.g., $\delta = 2$.

Applying this to the previous result, we observe

$$\hat{\theta}^{\text{aug}}_{\text{LCB}} - \tilde{\theta}_L = O_p(\min(\text{error}_n(\hat{\pi}), \text{error}_n(\hat{c}))).$$

However, the conditions of the theorem imply precisely that

$$|\theta_L - \tilde{\theta}_L| = \omega_p(\min(\text{error}_n(\hat{\pi}), \text{error}_n(\hat{c}))).$$

Thus, $\theta_L - \tilde{\theta}_L$ dominates $\hat{\theta}^{\text{aug}}_{\text{LCB}} - \tilde{\theta}_L$. Furthermore, weak duality implies that $\theta_L \geq \tilde{\theta}_L$ deterministically. Thus, using the decomposition

$$\theta_L - \hat{\theta}^{\text{aug}}_{\text{LCB}} = \underbrace{\theta_L - \tilde{\theta}_L}_{\text{strictly positive}} + \underbrace{\tilde{\theta}_L - \hat{\theta}^{\text{aug}}_{\text{LCB}}}_{\text{negligible}},$$

we conclude that $\liminf_{n\to\infty} \mathbb{P}(\hat{\theta}^{\text{aug}}_{\text{LCB}} \leq \theta_L) = 1$. As a result, under this form of misspecification, the dual bound $\hat{\theta}^{\text{aug}}_{\text{LCB}}$ is valid and in fact very conservative.

□

D.2 Technical lemmas for Section 3.1

Lemma D.1. *Assume the conditions of Theorem 3.5 except for "Condition 2." Then*

$$\sqrt{n}(\hat{\theta}^{\text{aug}}_{\text{LCB}} - \tilde{\theta}^{\text{aug}}_{\text{LCB}}) \xrightarrow{\text{p}} 0.$$

Proof. Recall that $\tilde{\theta}^{\text{aug}}_{\text{LCB}}, \tilde{\theta}^{\text{aug}}_{L}, \tilde{\sigma}^{\text{aug}}_{s}$ are defined exactly as $\hat{\theta}^{\text{aug}}_{\text{LCB}}, \hat{\theta}^{\text{aug}}_{L}, \sigma^{\text{aug}}_{s}$ are, but with $\hat{c}_1, \hat{c}_0, \hat{\pi}$ replaced with c_1, c_0 and π. Observe that

$$\sqrt{n}(\hat{\theta}^{\text{aug}}_{\text{LCB}} - \tilde{\theta}^{\text{aug}}_{\text{LCB}}) = \sqrt{n}(\hat{\theta}^{\text{aug}}_{L} - \tilde{\theta}^{\text{aug}}_{L}) + \sqrt{n}\Phi^{-1}(1-\alpha)\left(\frac{\tilde{\sigma}^{\text{aug}}_{s}}{\sqrt{n_2}} - \frac{\hat{\sigma}^{\text{aug}}_{s}}{\sqrt{n_2}}\right).$$

[11] Note that this analysis uses the condition that $\text{Var}(\tilde{S}_i \mid \mathcal{D}_1)$ is bounded away from zero, since otherwise a uniform CLT might not apply to $\tilde{\theta}^{\text{aug}}_{\text{LCB}}$—see the proof of Theorem 3.1 for details. Of course, this condition can be relaxed—see Remark 17.

By Slutsky's theorem, it suffices to show that $\sqrt{n}(\hat{\theta}_L^{\text{aug}} - \tilde{\theta}_L^{\text{aug}}) \xrightarrow{\text{p}} 0$, called "Claim 1," and that $\tilde{\sigma}_s^{\text{aug}} - \hat{\sigma}_s^{\text{aug}} \xrightarrow{\text{p}} 0$, called "Claim 2" (recall $n_2 \sim \frac{n}{2}$). We start by proving Claim 1.

Proof of Claim 1: We now show that $\sqrt{n}(\hat{\theta}_L^{\text{aug}} - \tilde{\theta}_L^{\text{aug}}) \xrightarrow{\text{p}} 0$. Our proof follows Wager (2020) with only minor adjustments. We note that by definition

$$\begin{aligned}\hat{\theta}_L^{\text{aug}} - \tilde{\theta}_L^{\text{aug}} = &\frac{1}{n_2}\sum_{i\in\mathcal{D}_2} \hat{c}_1(X_i) - c_1(X_i) + \frac{W_i}{\hat{\pi}(X_i)}(\hat{\nu}_1(Y_i, X_i) - \hat{c}_1(X_i)) - \frac{W_i}{\pi(X_i)}(\hat{\nu}_1(Y_i, X_i) - c_1(X_i)) \\ &+ \frac{1}{n_2}\sum_{i\in\mathcal{D}_2} \hat{c}_0(X_i) - c_0(X_i) + \frac{1-W_i}{1-\hat{\pi}(X_i)}(\hat{\nu}_0(Y_i, X_i) - \hat{c}_0(X_i)) - \frac{1-W_i}{1-\pi(X_i)}(\hat{\nu}_0(Y_i, X_i) - c_0(X_i)).\end{aligned}$$

The analysis of the two sums above is identical, so it suffices to show the first sum is $o_p(n^{-1/2})$. To do this, observe

$$\begin{aligned}&\frac{1}{n_2}\sum_{i\in\mathcal{D}_2} \hat{c}_1(X_i) - c_1(X_i) + \frac{W_i}{\hat{\pi}(X_i)}(\hat{\nu}_1(Y_i, X_i) - \hat{c}_1(X_i)) - \frac{W_i}{\pi(X_i)}(\hat{\nu}_1(Y_i, X_i) - c_1(X_i)) \\ =&\frac{1}{n_2}\sum_{i\in\mathcal{D}_2} (\hat{c}_1(X_i) - c_1(X_i))\left(1 - \frac{W_i}{\pi(X_i)}\right)\Bigg\} \text{ Term 1} \\ +&\frac{1}{n_2}\sum_{i\in\mathcal{D}_2} W_i(\hat{\nu}_1(Y_i, X_i) - c_1(X_i))\left(\hat{\pi}(X_i)^{-1} - \pi(X_i)^{-1}\right)\Bigg\} \text{ Term 2} \\ -&\frac{1}{n_2}\sum_{i\in\mathcal{D}_2} W_i(\hat{c}_1(X_i) - c_1(X_i))\left(\hat{\pi}(X_i)^{-1} - \pi(X_i)^{-1}\right)\Bigg\}. \text{ Term 3}\end{aligned}$$

To analyze these terms, we note the following.

1. For the first term, since $\pi(X_i)$ is a propensity score, we have that

$$\mathbb{E}\left[(\hat{c}_1(X_i) - c_1(X_i))\left(1 - \frac{W_i}{\pi(X_i)}\right) \mid \mathcal{D}_1, X_i\right] = 0$$

so these terms are mean zero conditional on $\mathcal{D}_1$. Furthermore,

$$\begin{aligned}&\text{Var}\left((\hat{c}_1(X_i) - c_1(X_i))\left(1 - \frac{W_i}{\pi(X_i)}\right) \mid \mathcal{D}_1\right) \\ =&\mathbb{E}\left[\text{Var}\left((\hat{c}_1(X_i) - c_1(X_i))\left(1 - \frac{W_i}{\pi(X_i)}\right) \mid \mathcal{D}_1, X_i\right) \mid \mathcal{D}_1\right] \\ =&\mathbb{E}\left[(\hat{c}_1(X_i) - c_1(X_i))^2\left(1 - \frac{1}{\pi(X_i)}\right)^2 \mid \mathcal{D}_1\right] \\ \leq&\left(1 - \frac{1}{\Gamma}\right)^2 \mathbb{E}[(\hat{c}_1(X_i) - c_1(X_i))^2 \mid \mathcal{D}_1] \\ =&O(1)\cdot \text{error}_n(\hat{c})^2.\end{aligned}$$

Thus, Chebyshev's inequality tells us that the first term is $O_p(\text{error}_n(\hat{c})^2/\sqrt{n}) = o_p(n^{-1/2})$ because we assume $\text{error}_n(\hat{c}) \xrightarrow{\text{p}} 0$.

2. For the second term, since $c_1(X_i) = \mathbb{E}[\hat{\nu}_1(Y_i, X_i) \mid \mathcal{D}_1, X_i]$, each of the summands are each mean zero conditional on X_i and $\mathcal{D}_1$. Their conditional variance is therefore

$$\begin{aligned}&\text{Var}\left(W_i(\hat{\nu}_1(Y_i, X_i) - c_1(X_i))\left(\hat{\pi}(X_i)^{-1} - \pi(X_i)^{-1}\right) \mid \mathcal{D}_1\right) \\ =&\mathbb{E}\left[\text{Var}\left(W_i(\hat{\nu}_1(Y_i, X_i) - c_1(X_i))\left(\hat{\pi}(X_i)^{-1} - \pi(X_i)^{-1}\right) \mid X_i, \mathcal{D}_1\right) \mid \mathcal{D}_1\right] \\ =&\mathbb{E}\left[\left(\hat{\pi}(X_i)^{-1} - \pi(X_i)^{-1}\right)^2 \mathbb{E}\left[W_i(\hat{\nu}_1(Y_i, X_i) - c_1(X_i))^2 \mid \mathcal{D}_1, X_i\right] \mid \mathcal{D}_1\right] \\ \leq&\mathbb{E}[\left(\hat{\pi}(X_i)^{-1} - \pi(X_i)^{-1}\right)^2 \mathbb{E}[\hat{\nu}_1(Y_i(1), X_i)^2 \mid \mathcal{D}_1, X_i] \mid \mathcal{D}_1] \\ \leq&\mathbb{E}[\left(\hat{\pi}(X_i)^{-1} - \pi(X_i)^{-1}\right)^{2s} \mid \mathcal{D}_1]^{1/s}\mathbb{E}[\hat{\nu}_1(Y_i(1), X_i)^{2+\delta} \mid \mathcal{D}_1]^{1/q} \qquad \text{taking, e.g., } \delta = 2 \\ =&o_P(1).\end{aligned}$$

where the penultimate inequality follows from Holder's inequality with $s, q \geq 1$ satisfying $\frac{1}{s} + \frac{1}{q} = 1, 2^q = 2+\delta$. The last equality follows from the fact $\mathbb{E}[\hat{\nu}_1(Y_i(1), X_i)^{2+\delta} \mid \mathcal{D}_1]$ is uniformly bounded and $\hat{\pi}$ and π are uniformly bounded; thus $\mathbb{E}[\left(\hat{\pi}(X_i)^{-1} - \pi(X_i)^{-1}\right)^{2s} \mid \mathcal{D}_1]^{1/s} = o_p(1)$ because $\text{error}_n(\hat{\pi}) \coloneqq \mathbb{E}[(\hat{\pi}(X_i) - \pi(X_i))^2 \mid \mathcal{D}_1] = o(1)$. As a result, Chebyshev's inequality implies that the second term is $O_p\left(\frac{o(1)}{\sqrt{n_2}}\right) = o_p(n^{-1/2})$.

3. For the third term, we merely apply Cauchy-Schwartz. In particular,

$$\begin{aligned}
&\mathbb{E}\left[\left|\frac{1}{n_2}\sum_{i\in\mathcal{D}_2} W_i(\hat{c}_1(X_i) - c_1(X_i))\left(\hat{\pi}(X_i)^{-1} - \pi(X_i)^{-1}\right)\right|\right] \\
\leq &\sqrt{\mathbb{E}\left[\frac{1}{n_2}\sum_{i\in\mathcal{D}_2}(\hat{c}_1(X_i) - c_1(X_i))^2\right] \cdot \mathbb{E}\left[\frac{1}{n_2}\sum_{i\in\mathcal{D}_2}(\hat{\pi}(X_i)^{-1} - \pi(X_i)^{-1})^2\right]} \\
\leq &O(1)\sqrt{\mathbb{E}[\text{error}_n(\hat{c})^2]\mathbb{E}[\text{error}_n(\hat{\pi})^2]} \\
= &o(n^{-1/2})
\end{aligned}$$

where the penultimate line follows from the fact that the terms in the sums of the expectations are i.i.d. conditional on $\mathcal{D}_1$, the definition of $\text{error}_n(\hat{c}), \text{error}_n(\hat{\pi})$, and the overlap assumptions on $\hat{\pi}$ and π. The last line follows from the assumption in Condition 2.

This shows that $\hat{\theta}_L^{\text{aug}} - \tilde{\theta}_L^{\text{aug}} = o_p(n^{-1/2})$, proving Claim 1.

Proof of Claim 2: We now show that $\hat{\sigma}_s^{\text{aug}} - \tilde{\sigma}_s^{\text{aug}} = o_p(1)$. To do this, define

$$\tilde{S}_i \coloneqq c_1(X_i) - c_0(X_i) + \frac{W_i}{\pi(X_i)}(\hat{\nu}_1(Y_i, X_i) - c_1(X_i)) + \frac{1-W_i}{1-\pi(X_i)}(\hat{\nu}_0(Y_i, X_i) - c_0(X_i)). \tag{82}$$

which is simply the definition of S_i but with $\hat{c}_1, \hat{c}_0, \hat{\pi}$ replaced with c_1, c_0, π. Observe that

$$(\hat{\sigma}_s^{\text{aug}})^2 = \frac{1}{n_2}\sum_{i\in\mathcal{D}_2} S_i^2 - (\hat{\theta}_L^{\text{aug}})^2 \text{ and } (\tilde{\sigma}_s^{\text{aug}})^2 = \frac{1}{n_2}\sum_{i\in\mathcal{D}_2} \tilde{S}_i^2 - (\tilde{\theta}_L^{\text{aug}})^2.$$

We already showed that $\hat{\theta}_L^{\text{aug}} - \tilde{\theta}_L^{\text{aug}} = o_p(1)$, so it suffices to show that

$$\frac{1}{n_2}\sum_{i\in\mathcal{D}_2} S_i^2 - \tilde{S}_i^2 = o_p(1).$$

To do this, it suffices to show that $\mathbb{E}[|S_i^2 - \tilde{S}_i^2|] = o(1)$ for any $i \in \mathcal{D}_2$. Note

$$|S_i^2 - \tilde{S}_i^2| \leq |S_i + \tilde{S}_i||S_i - \tilde{S}_i| \implies \mathbb{E}[|S_i^2 - \tilde{S}_i^2|] \leq \sqrt{\mathbb{E}[(S_i + \tilde{S}_i)^2]\mathbb{E}[(S_i - \tilde{S}_i)^2]}$$

and the moment conditions in the theorem imply that $\mathbb{E}[(S_i + \tilde{S}_i)^2]$ is uniformly bounded. Therefore, it suffices to show that $S_i - \tilde{S}_i = o_{L_2}(1)$. However, we already showed this in the proof of Claim 1 (combining the analysis of all three terms), when we showed that conditional on $\mathcal{D}_1$, $|S_i - \tilde{S}_i|$ has mean of order $o(n^{-1/2})$ and its variance conditional on $\mathcal{D}_1$ is bounded by $\max(\text{error}_n(\hat{\pi})^2, \text{error}_n(\hat{c})^2)$, which is $o_{L_2}(1)$ by the assumption in the theorem. This concludes the proof. □

Lemma D.2. *Assume the conditions of Theorem 3.5 except for "Condition 1." Then*

$$\tilde{\theta}_L - \hat{\theta}_{\text{LCB}}^{\text{aug}} = O_p(n^{-1/2}) + O_p(\min(\text{error}_n(\hat{\pi}), \text{error}_n(\hat{c}))).$$

Proof. We will first show that $\tilde{\theta}_L - \hat{\theta}_L^{\text{aug}} = o_P(1)$. We begin by writing

$$\hat{\theta}_L^{\text{aug}} \coloneqq \frac{1}{n_2}\sum_{i\in\mathcal{D}_2} S_i$$

where

$$S_i \coloneqq \hat{c}_1(X_i) + \hat{c}_0(X_i) + \frac{W_i}{\hat{\pi}(X_i)}(\hat{\nu}_1(Y_i, X_i) - \hat{c}_1(X_i)) + \frac{1-W_i}{1-\hat{\pi}(X_i)}(\hat{\nu}_0(Y_i, X_i) - \hat{c}_0(X_i)).$$

To show this initial result, there are two cases. Throughout the proof, as notation, let $Y(1), Y(0), X \in \mathbb{R}^p$ be an i.i.d. draw from the law of $Y(1), Y(0), X$ which is independent of the data. Recall also that $n_2 \coloneqq |\mathcal{D}_2| \sim n/2$ by assumption.

Case 1: For this case, we can decompose

$$\begin{aligned}
\hat{\theta}_L^{\text{aug}} &= \frac{1}{n_2} \sum_{i \in \mathcal{D}_2} c_1(X_i) + c_0(X_i) \\
&+ \frac{1}{n_2} \sum_{i \in \mathcal{D}_2} \frac{W_i}{\hat{\pi}(X_i)} (\hat{\nu}_1(Y_i, X_i) - c_1(X_i)) + \frac{1 - W_i}{1 - \hat{\pi}(X_i)} (\hat{\nu}_0(Y_i, X_i) - c_0(X_i)) \\
&+ \frac{1}{n_2} \sum_{i \in \mathcal{D}_2} \hat{c}_1(X_i) - c_1(X_i) + \hat{c}_0(X_i) - c_0(X_i) \\
&+ \frac{1}{n_2} \sum_{i \in \mathcal{D}_2} \frac{W_i}{\hat{\pi}(X_i)} (\hat{c}_1(Y_i, X_i) - c_1(X_i)) + \frac{1 - W_i}{1 - \hat{\pi}(X_i)} (\hat{c}_0(Y_i, X_i) - c_0(X_i)).
\end{aligned}$$

We now analyze these terms in order.

1. For the first term, note that e.g. for $\delta = 0$, $\mathbb{E}[|c_k(X)|^{2+\delta} \mid \mathcal{D}_1] \leq \mathbb{E}[|\hat{\nu}_k(Y(k), X)|^{2+\delta} \mid \mathcal{D}_1] \leq B$ is uniformly bounded by Assumption 3.2. As a result, Chebyshev's inequality implies that

$$\frac{1}{n_2} \sum_{i \in \mathcal{D}_2} c_1(X_i) - c_0(X_i) - \mathbb{E}[c_1(X) - c_0(X) \mid \mathcal{D}_1] = \frac{1}{n_2} \sum_{i \in \mathcal{D}_2} c_1(X_i) - c_0(X_i) - \tilde{\theta}_L = O_p(n_2^{-1/2}).$$

   Therefore, it suffices to show that the following terms vanish in probability.

2. The second term vanishes because $\mathbb{E}[\hat{\nu}_k(Y, X) - c_k(X)) \mid \mathcal{D}_1, X] = 0$ for $k \in \{0, 1\}$ by definition of c_1, c_0. Thus, the summands in the second term are all mean zero. Furthermore, for (e.g.) $\delta = 0$, their $2 + \delta$ moment is uniformly bounded by Assumption 3.2 and the fact that $\hat{\pi}(X_i) \in (\Gamma, 1 - \Gamma)$ for some $\Gamma > 0$. Thus, Chebyshev's inequality implies that this term is $O_p(n^{-1/2})$.

3. The third term vanishes in probability because we assume that the conditional $2 + \delta$ moment of $\hat{c}_1(X), \hat{c}_0(X)$ is uniformly bounded, and this is also true for $c_0(X), c_1(X)$ by Assumption 3.2. Thus Chebysehv's inequality tells us that

$$\frac{1}{n_2} \sum_{i \in \mathcal{D}_2} \hat{c}_1(X_i) - c_1(X_i) - \hat{c}_0(X_i) + c_0(X_i) = O_p(\mathbb{E}[|\hat{c}_1(X) - c_1(X) + \hat{c}_0(X) - c_0(X)| \mid \mathcal{D}_1] + n_2^{-1/2}).$$

   We also observe that by the triangle inequality and Jensen's inequality,

$$\begin{aligned}
\mathbb{E}[|\hat{c}_1(X) - c_1(X) + \hat{c}_0(X) - c_0(X)| \mid \mathcal{D}_1] &\leq \sum_{k \in \{0,1\}} \mathbb{E}[|\hat{c}_k(X) - c_k(X)| \mid \mathcal{D}_1] \\
&\leq 2 \max_{k \in \{0,1\}} \mathbb{E}[(\hat{c}_k(X) - c_k(X))^2 \mid \mathcal{D}_1]^{1/2} \\
&= 2 \cdot \text{error}_n(\hat{c}).
\end{aligned}$$

   Thus, the third term is $O_p(\text{error}_n(\hat{c}) + n^{-1/2})$.

4. The fourth term vanishes in probability by the same argument as the second term.

Combining these results shows that $\hat{\theta}_L^{\text{aug}} - \tilde{\theta}_L = O_p(\text{error}_n(\hat{c}) + n^{-1/2})$ by assumption.

Case 2: In this case, we assume $\mathbb{E}[|\hat{\pi}(X) - \pi(X)| \mid \mathcal{D}_1] = o_p(1)$. We can decompose

$$\begin{aligned}
\hat{\theta}_L^{\text{aug}} &:= \frac{1}{n_2}\sum_{i\in\mathcal{D}_2} \hat{c}_1(X_i) + \hat{c}_0(X_i) + \frac{W_i}{\hat{\pi}(X_i)}(\hat{\nu}_1(Y_i, X_i) - \hat{c}_1(X_i)) + \frac{1-W_i}{1-\hat{\pi}(X_i)}(\hat{\nu}_0(Y_i, X_i) - \hat{c}_0(X_i)) \\
&= \frac{1}{n_2}\sum_{i\in\mathcal{D}_2} \hat{c}_1(X_i)\left[1 - \frac{W_i}{\pi(X_i)}\right] + \hat{c}_0(X_i)\left[1 - \frac{1-W_i}{1-\pi(X_i)}\right]\Bigg\} \text{ term 1} \\
&+ \frac{1}{n_2}\sum_{i\in\mathcal{D}_2} \frac{W_i\hat{\nu}_1(Y_i, X_i)}{\pi(X_i)} + \frac{(1-W_i)\hat{\nu}_0(Y_i, X_i)}{1-\pi(X_i)}\Bigg\} \text{ term 2} \\
&+ \frac{1}{n_2}\sum_{i\in\mathcal{D}_2} \left(\hat{\pi}(X_i)^{-1} - \pi(X_i)^{-1}\right) W_i(\hat{\nu}_1(Y_i, X_i) - \hat{c}_1(X_i))\Bigg\} \text{ term 3} \\
&+ \frac{1}{n_2}\sum_{i\in\mathcal{D}_2} ((1-\hat{\pi}(X_i))^{-1} - (1-\pi(X_i))^{-1})(1-W_i)(\hat{\nu}_0(Y_i, X_i) - \hat{c}_0(X_i)\Bigg\} \text{ term 4}
\end{aligned}$$

To analyze these terms, observe that

1. The summands in term 1 are mean zero and i.i.d. conditional on $\mathcal{D}_1$. Furthermore, their $2+\delta$ moment is uniformly bounded conditional on $\mathcal{D}_1$ since the $2+\delta$ moments of $\hat{c}_1, \hat{c}_0$ are uniformly bounded and $\pi(X_i)$ is bounded away from zero and one. Chebyshev's inequality thus implies that Term 1 is $O_p(n^{-1/2})$.

2. Term 2 is simply an IPW estimator of $\tilde{\theta}_L$, so under the assumptions of Theorem 3.1 it converges to $\tilde{\theta}_L$ plus $O_p(n^{-1/2})$.

3. Term 3 is an i.i.d. sum conditional on $\mathcal{D}_1$. Its summands have uniformly bounded $2+\delta$ moment because Assumption 3.2 implies that $\hat{\nu}_1(Y(1), X) - \hat{c}_1(X)$ has a uniformly bouned $2+\delta$th moment and $\hat{\pi}(X_i)^{-1}, \pi(X_i)^{-1}$ are uniformly bounded. Furthermore, Holder's inequality yields that

$$\begin{aligned}
&\mathbb{E}\left[\left|\left(\hat{\pi}(X_i)^{-1} - \pi(X_i)^{-1}\right)(W_i(\hat{\nu}_1(Y_i, X_i) - \hat{c}_1(X_i)))\right| \mid \mathcal{D}_1\right] \\
\leq &\sqrt{\mathbb{E}[(\hat{\pi}(X_i)^{-1} - \pi(X_i)^{-1})^2 \mid \mathcal{D}_1]\mathbb{E}\left[(W_i(\hat{\nu}_1(Y_i, X_i) - \hat{c}_1(X_i))^2 \mid \mathcal{D}_1\right]} \\
= &O\left(\sqrt{\mathbb{E}[(\hat{\pi}(X_i) - \pi(X_i))^2 \mid \mathcal{D}_1]}\right) \\
:= &O(\text{error}_n(\hat{\pi})).
\end{aligned}$$

   where penultimate line follows because (a) $\mathbb{E}\left[(W_i(\hat{\nu}_1(Y_i, X_i) - \hat{c}_1(X_i)))^2 \mid \mathcal{D}_1\right]$ is uniformly bounded and (b) $\hat{\pi}, \pi$ are uniformly bounded away from zero, and the function $1 \mapsto x^{-2}$ is Lipschitz on $[\Gamma, \infty)$. Thus, $\hat{\pi}(X_i)^{-1} - \pi(X_i)^{-1} \leq L(\hat{\pi}(X_i) - \pi(X_i))$ where L is some Lipschitz constant not depending on n.

   Applying Chebyshev's inequality to Term 3 yields that Term 3 is $O_p(\text{error}_n(\hat{\pi})) + O_p(n^{-1/2})$.

4. Term 4 is $O_p(\text{error}_n(\hat{\pi})) + O_p(n^{-1/2})$ for the same reason that Term 3 is.

Combining this analysis yields that $\hat{\theta}_L^{\text{aug}} = \tilde{\theta}_L + O_p(n^{-1/2}) + O(\text{error}_n(\hat{\pi}))$.

Combining cases 1 and 2 implies that

$$\hat{\theta}_L^{\text{aug}} - \tilde{\theta}_L = O_p(n^{-1/2}) + O_p(\min(\text{error}_n(\hat{\pi}), \text{error}_n(\hat{c}))).$$

To complete the proof of the lemma, we note that

$$\hat{\theta}_{\text{LCB}}^{\text{aug}} - \tilde{\theta}_L = \hat{\theta}_L^{\text{aug}} - \tilde{\theta}_L - \Phi^{-1}(1-\alpha)\frac{\hat{\sigma}_s^{\text{aug}}}{\sqrt{n_2}}.$$

Thus, to prove the lemma, it suffices to prove that $\hat{\sigma}_s^{\text{aug}} = O_p(1)$. To do this, we observe that $\hat{\sigma}_s$ is defined as

$$(\hat{\sigma}_s^{\text{aug}})^2 = \frac{1}{n_2}\sum_{i\in\mathcal{D}_2}(S_i - \bar{S})^2.$$

Since S_i is i.i.d. has a uniformly bounded $2+\delta$ moment conditional on $\mathcal{D}_1$, a uniform law of large numbers applied to S_i and then S_i^2 yields that $\hat{\sigma}_s^{\text{aug}} = \sqrt{\text{Var}(S_i \mid \mathcal{D}_1)} + o_p(1)$ where $\text{Var}(S_i \mid \mathcal{D}_1)$ is uniformly bounded by the theorem assumptions. This proves that $\hat{\sigma}_s^{\text{aug}} = O_p(1)$. □

D.3 Main proofs from Section 3.2

Theorem 3.2. *Suppose strong duality holds, i.e., $g(\nu^\star) = \theta_L$. Then*

$$0 \le \theta_L - \tilde{\theta}_L \le \mathbb{E}_{X \sim P_X^\star}\left[\text{error}_P(X) \cdot \text{error}_\nu(X)\right]. \tag{83}$$

Proof. As notation, for any functions $f_0, f_1, h_0, h_1 : \mathcal{Y} \times \mathcal{X} \to \mathbb{R}$ and any $x \in \mathcal{X}$, define the inner product

$$\langle (f_0, f_1), (h_0, h_1) \rangle_x = \sum_{k \in \{0,1\}} \int_{y \in \mathcal{Y}} f_k(y, x) h_k(y, x) \psi_x(dy).$$

Furthermore, let $\|(f_0, f_1)\|_X = \sqrt{\langle (f_0, f_1), (f_0, f_1) \rangle}$ denote the standard norm with respect to $\langle \cdot, \cdot \rangle_X$. It may be helpful to note that the definitions in Section 3.2 imply that

$$\text{error}_P(x) = \|\hat{p} - p^\star\|_X \text{ and } \text{error}_\nu(x) = \|\hat{\nu} - \nu^\star\|_X \tag{84}$$

where $\hat{p} = (\hat{p}_0, \hat{p}_1) : \mathcal{Y} \times \mathcal{X} \to \mathbb{R}^2$ denotes the estimated conditional densities of $Y(k) \mid X, k \in \{0, 1\}$ and $p^\star = (p_0^\star, p_1^\star) : \mathcal{Y} \times \mathcal{X} \to \mathbb{R}^2$ denote the true conditional densities. With this notation, let $\hat{g} : \mathcal{V} \to \mathbb{R}$ denote the estimate of the dual which plugs in $\hat{p}$ for $p^\star$:

$$\begin{aligned} \hat{g}(\nu) &\coloneqq \mathbb{E}_{\hat{P}_{Y(0)|X} \times P_X^\star}[\nu_0(Y(0), X)] + \mathbb{E}_{\hat{P}_{Y(1)|X} \times P_X^\star}[\nu_1(Y(1), X] \\ &= \mathbb{E}_{X \sim P_X^\star}\left[\langle \hat{p}, \nu \rangle_X\right]. \end{aligned}$$

With this notation, we observe that

$$\begin{aligned} \theta_L - \tilde{\theta}_L &= g(\nu^\star) - g(\hat{\nu}) && \text{by strong duality and defn. of } \tilde{\theta}_L \\ &\le g(\nu^\star) - \hat{g}(\nu^\star) + \hat{g}(\hat{\nu}) - g(\hat{\nu}) && \text{since } \hat{\nu} \coloneqq \arg\max_\nu \hat{g}(\nu) \\ &= \mathbb{E}_{X \sim P_X^\star}\left[\langle p^\star - \hat{p}, \nu^\star - \hat{\nu} \rangle_X\right] && \text{using linearity of inner products} \\ &\le \mathbb{E}_{X \sim P_X^\star}\left[\|p^\star - \hat{p}\|_X \|\nu^\star - \hat{\nu}\|_X\right] && \text{by Cauchy-Schwartz} \\ &= \mathbb{E}_{X \sim P_X^\star}\left[\text{error}_P(X) \cdot \text{error}_\nu(X)\right] && \text{by definition.} \end{aligned}$$

It may be helpful to note that the third-to-last equation uses the fact that for any fixed ν, $\hat{g}(\nu) - g(\nu) = \mathbb{E}_{X \sim P_X^\star}[\langle p - \hat{p}, \nu \rangle_X]$. This completes the proof. □

Lemma 3.3. *Suppose $\mathcal{Y}$ is finite and consider estimated dual variables $\hat{\nu}$ as defined in Eq. (12). There exist a collection of finite deterministic Lipschitz constants $\{H(x) : x \in \mathcal{X}\}$ depending only on $P^\star, \mathcal{P}$ and f such that for all $x \in \mathcal{X}$, there exist optimal dual variables $\nu^\star$ such that the following holds deterministically:*

$$\text{error}_\nu(x)^2 \coloneqq \sum_{k \in \{0,1\}} \sum_{y \in \mathcal{Y}} (\hat{\nu}_k(y, x) - \nu^\star(y, x))^2 \le H(x) \cdot \text{error}_P(x)^2.$$

Proof. We begin by introducing notation. Suppose $\mathcal{Y} = \{y_1, \ldots, y_m\}$ which is finite by assumption. For any dual variables $\nu_0, \nu_1 : \mathcal{Y} \times \mathcal{X} \to \mathbb{R}$ and $k \in \{0, 1\}$, let

$$\nu_{k,x} = [\nu_k(y_1, x), \ldots, \nu_k(y_m, x)] \in \mathbb{R}^m$$

denote the vector of values of ν_k with and let $\nu_x = [\nu_{0,x}, \nu_{1,x}] \in \mathbb{R}^{2m}$ denote the concatenation of $\nu_{0,x}, \nu_{1,x}$. Furthermore let $p_{k,i}^\star(x) = \mathbb{P}(Y(k) = y_i \mid X = x)$ denote the probability mass function of $Y(k) \mid X$ for $k \in \{0, 1\}, i \in [m]$. Let $p^\star(x) = [p_{0,1}^\star(x), \ldots, p_{0,m}^\star(x), p_{1,1}^\star(x), \ldots, p_{1,m}^\star(x)] \in \mathbb{R}^{2m}$ denote the concatenation of the PMFs of $Y(1)$ and $Y(0)$; furthermore, let $\hat{p}(x)$ denote the estimated version of this vector based on the estimated laws $\hat{P}_{Y(0)|X=x}, \hat{P}_{Y(1)|X=x}$. Note that under this notation, we have that

$$\text{error}_P(x)^2 = \|\hat{p}(x) - p^\star(x)\|_2^2 \text{ and } \text{error}_\nu(x)^2 \coloneqq \sum_{k \in \{0,1\}} \sum_{y \in \mathcal{Y}} (\hat{\nu}_k(y, x) - \nu^\star(y, x))^2 = \|\hat{\nu}_x - \nu_x^\star\|_2^2.$$

Thus, it suffices to show that there exists a universal constant $H(x)$ depending only on population quantities such that $\|\hat{\nu}_x - \nu_x^\star\|_2^2 \le H(x)\|\hat{p}(x) - p^\star(x)\|_2^2$. To do this, recall that in Section 4.2, we prove $\nu^\star \in$

$\arg\max_{\nu \in \mathcal{V}} g(\nu)$ is an optimal dual variable if the following holds for all x:

$$\begin{aligned}\nu_x^\star \in \underset{\nu_x \in \mathbb{R}^{2m}, \lambda_{x,1}, \ldots, \lambda_{x,L}}{\arg\max} \quad & \nu_x^T p^\star(x) \\ \text{s.t. } \nu_{0,x,j} + \nu_{1,x,i} + \sum_{\ell=1}^{L} \lambda_{x,\ell} w_{x,\ell}(y_i, y_j) \le f(y_j, y_i, x) \text{ for all } & i, j \in [m] \\ \lambda_{x,1}, \ldots, \lambda_{x,\ell} \ge 0.\end{aligned}$$

This is a finite-dimensional linear program; in particular, we can put this in a standard form by writing $\lambda_x = (\lambda_{x,1}, \ldots, \lambda_{x,L})$ and observing

$$\lambda_x^\star, \nu_x^\star \in \underset{\nu_x, \lambda_x}{\arg\max}\, \nu_x^T p^\star(x) \text{ s.t. } A \begin{bmatrix} \nu_x \\ \lambda_x \end{bmatrix} \le b$$

for some known deterministic matrix $A \in \mathbb{R}^{m^2+\ell}$ and constraint vector $b \in \mathbb{R}^{m^2+\ell}$. Furthermore, by definition, $\hat{\nu}_x$ is the (minimum norm) solution to the same problem which replaces $p^\star(x)$ with $\hat{p}(x)$:

$$\hat{\lambda}_x, \hat{\nu}_x \in \underset{\nu_x, \lambda_x}{\arg\max}\, \nu_x^T \hat{p}(x) \text{ s.t. } A \begin{bmatrix} \nu_x \\ \lambda_x \end{bmatrix} \le b$$

Thus, the relationship between $\|\nu_x^\star - \hat{\nu}_x\|_2$ and $\|p^\star(x) - \hat{p}(x)\|_2$ is related to the stability of linear programs; in particular, we leverage the theory of Hoffman constants (Hoffman, 1952; Robinson, 1973). We give a detailed review of this theory in Appendix D.4. The upshot is that Lemma D.6 (proved in Appendix D.4) implies that for any linear programs of the form above, we have that there exists some $H(x) < \infty$ such that $\|\nu_x^\star - \hat{\nu}_x\|_2^2 \le H(x)\|p^\star(x) - \hat{p}(x)\|_2^2$, where $H(x)$ is a finite constant depending only on $p^\star(x)$, A, and b. These are all population quantities, so this completes the proof. □

Remark 18 ($H(x)$ is "dimension-free")**.** In the most important special case where $f(y_0, y_1, x) = f(y_0, y_1)$ does not depend on x and $\mathcal{P}$ is the unconstrained set of all distributions over $\mathcal{Y}^2 \times \mathcal{X}$, then the matrix A and constraint b in the previous proof do not depend on x. As a result, $H(x)$ depends only on the conditional PMF of $Y(k) \mid X = x$ for $k \in \{0, 1\}$ and does not explicitly depend on the dimension of $X \in \mathbb{R}^p$.

This suggests that we should not expect $H(x)$ to grow with the dimension of X (although it may grow as the size of $\mathcal{Y}$, the support of Y, increases). Indeed, in many typical high-dimensional asymptotic regimes, the law of the conditional PMF of $Y(k) \mid X$ does not change with $X \in \mathbb{R}^p$. For example, consider a single-index model where $Y(k)$ only depends on X through a linear function $a_k^T X$, formally written as $Y(k) \perp\!\!\!\perp X \mid a_k^T X$ for some $a_k \in \mathbb{R}^p, k \in \{0, 1\}$. In this setting, the law of the conditional PMF $\{\mathbb{P}(Y(k) = y \mid X)\}_{y \in \mathcal{Y}}$ does not change with dimension as long as the laws of $a_k^T X$ do not change with n or p. For example, if $X \sim \mathcal{N}(0, \Sigma)$, this holds as long as the aggregate signal strength $a_k^T \Sigma a_k$ stays constant. Indeed, this condition exactly matches ones used to analyze (e.g.) high-dimensional linear and logistic regression (Sur and Candès, 2019); otherwise, in the case of linear regression, the variance of Y might diverge as $n, p \to \infty$. In such regimes, the law of $H(X)$ will not change even as the dimension of X grows arbitrarily.

Theorem 3.4. *Suppose Y has finite support $\mathcal{Y}$, $\theta(P) = \mathbb{E}_P[f(Y_i(0), Y_i(1), X_i)]$ can be written as an expectation, and $\mathcal{P}$ denotes the set of all probability distributions on $\mathcal{Y}^2 \times \mathcal{X}$. Suppose furthermore that the random Lipschitz constant $H(X)$ has two moments, i.e., $\mathbb{E}[|H(X)|^2] < \infty$.*

Finally, assume that $\text{error}_P(X) = o_{L_4}(n^{-1/4})$ as $n \to \infty$, where X denotes a fresh sample of covariates. Then there exist a sequence of optimal dual variables $\nu^{\star(n)} \in \arg\max_{\nu \in \mathcal{V}} g(\nu)$ such that if $\hat{\theta}_L^\star$, $\hat{\theta}_{\text{LCB}}^\star$ denote the oracle estimators which use $\nu^{\star(n)}$ in place of $\hat{\nu}$,

$$\sqrt{n}(\hat{\theta}_L - \hat{\theta}_L^\star) \xrightarrow{\text{p}} 0 \text{ and } \sqrt{n}(\hat{\theta}_{\text{LCB}} - \hat{\theta}_{\text{LCB}}^\star) \xrightarrow{\text{p}} 0. \tag{85}$$

Proof. First, we prove that $\sqrt{n}(\hat{\theta}_L - \hat{\theta}_L^\star) \xrightarrow{\text{p}} 0$. To do this, note by definition

$$\hat{\theta}_L - \hat{\theta}_L^\star := \frac{1}{n_2} \sum_{i \in \mathcal{D}_2} \frac{W_i}{\pi(X_i)} \left(\hat{\nu}_1(Y_i(1), X_i) - \nu_1^{\star(n)}(X_i, Y_i(1)) \right) + \frac{1 - W_i}{1 - \pi(X_i)} \left(\hat{\nu}_0(Y_i(0), X_i) - \nu_0^{\star(n)}(X_i, Y_i(0)) \right).$$

Call each term in the conditionally i.i.d. sum above M_i. To show the sum above is $o_p(n^{-1/2})$, we will show that $\mathbb{E}[M_i \mid \mathcal{D}_1] = o_p(n^{-1/2})$ and $\text{Var}(M_i \mid \mathcal{D}_1) = o_p(1)$, at which point Chebyshev's inequality will imply

that the above sum is $o_p(n^{-1/2})$ asymptotically. For all steps, the starting point is Lemma 3.3, which says that there exist dual variables $\nu^{\star(n)} \in \arg\max_{\nu \in \mathcal{V}} g(\nu)$ satisfying

$$\max_{k \in \{0,1\}} \max_{y \in \mathcal{Y}} (\nu_k^{\star(n)}(y, x) - \hat{\nu}_k(y, x))^2 \leq \text{error}_\nu(x)^2 \leq H(x)\text{error}_P(x)^2,$$

for a set of deterministic Hoffman constants $\{H(x) : x \in \mathcal{X}\}$.

Step 1: analyzing the mean. Since $\pi(X_i)$ are the true propensity scores and $\hat{\nu}_1, \hat{\nu}_0$ are estimated on $\mathcal{D}_1$ which is independent of $\mathcal{D}_2$, we have that

$$\begin{aligned}
\mathbb{E}[M_i \mid \mathcal{D}_1] &= \mathbb{E}[\hat{\nu}_1(X, Y(1)) + \hat{\nu}_0(X, Y(0)) \mid \mathcal{D}_1] - \mathbb{E}[\nu_1^{\star(n)}(X, Y(1)) + \nu_0^{\star(n)}(X, Y(0))] \\
&:= \tilde{\theta}_L - \theta_L \\
&\leq \mathbb{E}_{X \sim P_X^\star}[\text{error}_P(X) \cdot \text{error}_\nu(X)] && \text{by Theorem 3.2} \\
&\leq \mathbb{E}_{X \sim P_X^\star}[\text{error}_P(X)^2 \sqrt{H(X)}] && \text{by Lemma 3.3} \\
&\leq \sqrt{\mathbb{E}_{X \sim P_X^\star}[\text{error}_P(X)^4]\mathbb{E}[|H(X)|]} \\
&= o_p(n^{-1/2}).
\end{aligned}$$

where the last step uses the assumptions that $\mathbb{E}[|H(X)|] < \infty$ and $\text{error}_P(X) = o_{L_4}(n^{-1/4})$.

Step 2: analyzing the variance. Lemma 3.3 yields that

$$\sum_{k \in \{0,1\}} (\hat{\nu}_k(X_i, Y_i(k)) - \nu_k^{\star(n)}(X_i, Y_i(k)))^2 \leq H(X_i)\text{error}_P(X_i)^2.$$

Therefore,

$$M_i^2 \leq \frac{H(X_i)\text{error}_P(X_i)^2}{\min(\pi(X_i), 1 - \pi(X_i))^2} \leq \Gamma^{-2} H(X_i)\text{error}_P(X_i)^2$$

where the last inequality follows by the strict overlap assumption that $\pi(X_i) \in [\Gamma, 1 - \Gamma]$ for some $\Gamma > 0$. From this, we apply Holder's inequality to conclude

$$\begin{aligned}
\text{Var}(M_i^2 \mid \mathcal{D}_1) &\leq \mathbb{E}[M_i^2 \mid \mathcal{D}_1] \\
&\leq \Gamma^{-2} \sqrt{\mathbb{E}[H(X_i)^2 \mid \mathcal{D}_1]\mathbb{E}[\text{error}_P(X_i)^4 \mid \mathcal{D}_1]} \\
&= o_p(1)\mathbb{E}[H(X_i)^2]
\end{aligned}$$

where the last line follows because (i) $H(X_i)$ is independent of $\mathcal{D}_1$ and (ii) $\text{error}_P(X_i) = o_p(1)$ by assumption, and furthermore $\text{error}_P(X_i)$ is uniformly bounded since it is the concatenation of two differences between probability vectors. Thus, $\text{Var}(M_i^2 \mid \mathcal{D}_1) = o_p(1)$ because $\mathbb{E}[H(X_i)^2] < \infty$ by assumption.

As a result, Chebyshev's inequality applied conditionally on $\mathcal{D}_1$ implies that

$$\hat{\theta}_L - \hat{\theta}_L^\star = \frac{1}{n_2} \sum_{i \in \mathcal{D}_2} M_i = O_p(|\mathbb{E}[M_i \mid \mathcal{D}_1]| + \sqrt{\text{Var}(M_i \mid \mathcal{D}_1)}/\sqrt{n_2}) = o_p(n^{-1/2})$$

which completes the first part of the proof.

Now, we show that $\sqrt{n}(\hat{\theta}_{\text{LCB}} - \hat{\theta}_{\text{LCB}}^\star) \xrightarrow{\text{p}} 0$. As notation, let $S_i = \frac{W_i}{\pi(X_i)}\hat{\nu}_1(Y_i(1), X_i) + \frac{1-W_i}{1-\pi(X_i)}\hat{\nu}_0(Y_i(0), X_i)$ and let $S_i^\star$ denote the same quantity with $\nu^{\star(n)}$ replaced with ν. If $\hat{\sigma}_s$ and $\hat{\sigma}_s^\star$ are the empirical standard deviations of $\{S_i\}_{i \in \mathcal{D}_2}$ and $\{S_i^\star\}_{i \in \mathcal{D}_2}$, respectively, then by definition

$$\hat{\theta}_{\text{LCB}} - \hat{\theta}_{\text{LCB}}^\star = \hat{\theta}_L - \hat{\theta}_L^\star + \frac{\Phi^{-1}(1 - \alpha)}{\sqrt{n_2}} (\hat{\sigma}_s - \hat{\sigma}_s^\star).$$

Therefore, to complete the proof, it suffices to show that $\hat{\sigma}_s - \hat{\sigma}_s^\star \xrightarrow{\text{p}} 0$. However, we already showed that if $M_i = S_i - S_i^\star$, then

$$\mathbb{E}[M_i^2 \mid \mathcal{D}_1] = o_p(1).$$

In other words, $\hat{\sigma}_s$ and $\hat{\sigma}_s^\star$ are the sample standard deviations of i.i.d. summands whose difference vanishes as $n \to \infty$. Furthermore, each summand (whether S_i or $S_i^\star$) has uniformly bounded $2 + \delta$ moments as well. Thus, the exact same argument as in Lemma D.1 shows that $\hat{\sigma}_s - \hat{\sigma}_s^\star \xrightarrow{\text{p}} 0$, proving the result. □

Remark 19. Theorem 3.4 assumes that $H(X)$ has a second moment. It is possible to generalize beyond this as long as $H(X)\text{error}_P(X)^2 = o_{L_3}(1)$ and $H(X)\text{error}_P(X)^2 = o_{L_1}(n^{-1/2})$, in which case the proof will go through without modification.

Finally, we prove Proposition D.2, which proves that under certain conditions, if $\sup_{x\in\mathcal{X}} \text{error}_P(x)$ lies below a certain threshold, then $\hat{\nu} = \nu^\star$ deterministically. As in Section 3.2, we emphasize that Proposition D.2 is not too informative in practice, since (i) in finite samples, we do not expect $\sup_{x\in\mathcal{X}} \text{error}_P(x)$ to be small enough such that $\hat{\nu} = \nu$, and (ii) it relies on a series of technical regularity conditions which may be hard to interpret or verify. However, it is an interesting theoretical result nonetheless since it does not require $\sup_{x\in\mathcal{X}} \text{error}_P(x)$ to vanish at any particular rate.

Note that for this result, we assume that $\mathcal{P}$ is the unconstrained set of all distributions on $\mathcal{Y}^2 \times \mathcal{X}$. To state the additional assumptions of Proposition D.2, recall that for each $x \in \mathcal{X}$, the relevant dual variables $\nu^\star(\cdot, x)$ can be defined as the solution to a linear program of the form

$$\nu_x^\star = \arg\max_v p^\star(x)^T v \text{ s.t. } Av \le b(x)$$

for $A \in \mathbb{R}^{m^2 \times 2m}$, and $\hat{\nu}_x$ is the solution to the same problem but replacing $\hat{p}(x)$ with $p^\star(x)$. Now, the key observation leading to Proposition D.2 is that under certain conditions, if $\hat{p}(x)$ is close enough to $p^\star(x)$, then $\hat{\nu}_x = \nu_x^\star$ holds exactly. To quantify this relationship, we define $J(x)$ to be the set of constraints above which are satisfied with equality by the optimal value:

$$J(x) \coloneqq \{j \in [m^2] : A_j^T \nu_x^\star = b(x)\} \tag{86}$$

and we define $C(x)$ to be the cone formed by $\{-A_j^T : j \in J(x)\}$:

$$C(x) \coloneqq \left\{ \sum_{j\in J(x)} -\lambda_j A_j^T : \lambda_j \ge 0 \ \forall j \in J(x) \right\}. \tag{87}$$

As we will see in the proof of Proposition D.2, $p^\star(x) \in C(x)$ holds by the standard optimality conditions for linear programs. Roughly speaking, the condition we require for Proposition D.2 is that $p^\star(x)$ is in the interior of $C(x)$, and is uniformly bounded away from the boundary.

Formally, the condition we require is slightly weaker than that because we can leverage the fact that $p^\star(x)$ and $\hat{p}(x)$ are restricted to lie in $\Delta(m)^2$, where $\Delta(m)$ denotes the probability simplex on $\mathbb{R}^m$. Now, we define the *threshold* function $t(x)$ to be the distance between $p^\star(x)$ and the intersection of the exterior of $C(x)$ and $\Delta(m)^2$:

$$t(x) \coloneqq \inf_{u\in\Delta(m)^2, u\notin C(x)} \|p^\star(x) - u\|_\infty.$$

If $\sup_{x\in\mathcal{X}} \text{error}_P(x) \to 0$ and $t(x)$ is uniformly bounded away from zero, then $\hat{\theta}_L = \hat{\theta}_L^\star$ with probability one asymptotically, as formalized by the following proposition. These assumptions may or may not be plausible; again, we emphasize that this result is mainly a theoretical curiosity, and we do not think it is as practically informative as Theorem 3.4.

Proposition D.2. *Suppose $\mathcal{Y}$ has finite support, $\theta(P) = \mathbb{E}[f(Y_i(0), Y_i(1))]$ can be written as an expectation, and that $\hat{\nu} = \arg\max_{\nu\in\mathcal{V}} \hat{g}(\nu)$ is a.s. unique. Suppose that the first-stage error converges uniformly to zero, i.e., $\sup_{x\in\mathcal{X}} \text{error}_P(x) \xrightarrow{\text{p}} 0$ and that the threshold function $t(x) \ge 0$ is uniformly bounded away from zero. Then*

$$\liminf_n \mathbb{P}(\hat{\theta}_L = \hat{\theta}_L^\star) = 1 \text{ and } \liminf_n \mathbb{P}(\hat{\theta}_{\text{LCB}} = \hat{\theta}_{\text{LCB}}^\star) = 1.$$

Proof. Recall that for any $x \in \mathcal{X}$, we can write that

$$\nu_x^\star \in \arg\max_v p^\star(x)^T v \text{ s.t. } Av \le b(x)$$

for the $A, b(x)$ specified previously. For any dual variables $\lambda \in \mathbb{R}^{m^2}_{\ge 0}$, the KKT equations for this linear program are

$$\begin{aligned} A\nu_x^\star &\le b(x) && \text{primal feasibility} \\ \lambda &\ge 0 && \text{dual feasibility} \\ \lambda^T(A\nu_x^\star - b(x)) &= 0 && \text{complimentary slackness} \\ A^T\lambda + p^\star(x) &= 0 && \text{Lagrange condition} \end{aligned}$$

where the third inequality is equivalent to the restriction that $\lambda_j = 0$ for any $j \notin J(x)$ as defined in Eq. (86). As a result, the fourth inequality is equivalent to $A_{J(x)}^T \lambda_{J(x)} + p^\star(x) = 0$ which is equivalent to $p^\star(x) \in C(x)$ by definition of $C(x)$ as in Eq. (87). Overall, this implies that if we replace $p^\star(x)$ with $\hat{p}(x)$, $\nu_x^\star$ remains primal optimal as long as $\hat{p}(x) \in C(x)$.

Since $\hat{p}(x)$ is a finite dimensional vector, we note that $\|\hat{p}(x) - p^\star(x)\|_\infty \leq \beta \|\hat{p}(x) - p^\star(x)\|_2 = \beta \text{error}_P(x)$ for some constant $\beta > 0$. By definition of $t(x)$, we have that $\|\hat{p}(x) - p^\star(x)\|_\infty \leq \beta \text{error}_P(x) \leq t(x)$ implies that $\hat{p}(x) \in C(x)$. Furthermore, we assumed that $\hat{\nu}_x$ is a.s. unique, and therefore $\hat{\nu}_x = \nu_x^\star$ whenever $\text{error}_P(x) \leq \frac{1}{\beta} t(x)$.

Note that $\hat{\theta}_L, \hat{\theta}_{\text{LCB}}$ only depend on $\hat{\nu}$ through $\{\hat{\nu}_{X_i}\}_{i \in \mathcal{D}_2}$ as per Section 4.1. Therefore,

$$\begin{aligned}
\mathbb{P}(\hat{\theta}_L = \hat{\theta}_L^\star \text{ and } \hat{\theta}_{\text{LCB}} = \hat{\theta}_{\text{LCB}}^\star) &\geq \mathbb{P}\left(\cap_{i \in \mathcal{D}_2} \hat{\nu}_{X_i} = \nu_{X_i}^\star\right) \\
&\geq \mathbb{P}\left(\cap_{i \in \mathcal{D}_2} \text{error}_P(X_i) \leq t(X_i)/\beta\right) \\
&\geq \mathbb{P}\left(\sup_{x \in \mathcal{X}} \text{error}_P(x) \leq \inf_{i \in \mathcal{D}_2} t(X_i)/\beta\right) \\
&\to 1
\end{aligned}$$

where the last step follows because we assume $\inf_{i \in \mathcal{D}_2} t(X_i)$ is bounded above some constant and that $\sup_{x \in \mathcal{X}} \text{error}_P(x) = 0$. □

D.4 Theory of Hoffman constants

In this section, we review the definition of a Hoffman constant and prove Lemma D.6, the key stability result for linear programs that underlies Lemma 3.3. First, we review what a Hoffman constant is.

Lemma D.5 (Hoffman constant)**.** *For matrices $A \in \mathbb{R}^{m \times n}, C \in \mathbb{R}^{k \times n}$, define the set*

$$M(b, d) = \{x \in \mathbb{R}^n : Ax \leq b, Cx = d\} \text{ for } b \in \mathbb{R}^m, d \in \mathbb{R}^k.$$

Hoffman (1952), Robinson (1973) showed that there exists a constant $H(A, C) < \infty$ such that for all $x \in \mathbb{R}^n, b \in \mathbb{R}^m, d \in \mathbb{R}^k$ s.t. $M(b, d) \neq \emptyset$, we have

$$\text{dist}(x, M(b, d)) \leq H(A, C) \left\| \begin{bmatrix} (Ax - b)_+ \\ Cx - d \end{bmatrix} \right\|_2. \tag{88}$$

We now prove the key technical lemma underlying Theorem 3.4, using ideas from Robinson (1973).

Lemma D.6 (Robinson (1973) Corollary 3.1)**.** *Consider the LP*

$$\min_x c^T x \text{ s.t. } Ax \leq b \tag{89}$$

whose dual is

$$\max_{y \geq 0} -b^T y \text{ s.t. } A^T y + c = 0. \tag{90}$$

Suppose the primal and the dual are both feasible. Suppose that $\hat{x} \in \mathbb{R}^n, \hat{y} \in \mathbb{R}^m$ are minimum norm solution solving (89), (90) but with $\hat{c}$ replacing c. Then there exists some optimal solution $x^\star \in \mathbb{R}^n, y^\star \in \mathbb{R}^m$ to (89), (90) such that

$$\|(\hat{x}, \hat{y}) - (x^\star, y^\star)\|_2 \leq \sigma \|c - \hat{c}\|_2$$

where σ is a scaled Hoffman constant that depends on A, b and c.

Proof. Strong duality holds for primal-dual feasible linear programs. Thus, by strong duality, a pair (x, y) is primal-dual optimal if and only if

$$\begin{aligned}
Ax \leq b &\text{ primal feasibility} \\
y \geq 0 &\text{ dual feasibility} \\
A^T y = c &\text{ dual feasibility} \\
c^T x^\star - b^T y^\star = 0 &\text{ dual gap is zero.}
\end{aligned}$$

We note that this "dual gap" condition is slightly different than the standard KKT conditions, e.g., from Boyd and Vandenberghe (2004); this innovation, due to Robinson (1973), is what allows us to apply Lemma D.5.[12]

With this characterization, for $z = (x, y) \in \mathbb{R}^{n+m}$, the optimality conditions are equivalent to the following:

$$A_0 z \leq \begin{bmatrix} b \\ 0 \end{bmatrix} \text{ and } C_0 z = \begin{bmatrix} c \\ 0 \end{bmatrix} \tag{91}$$

where

$$A_0 \coloneqq \begin{bmatrix} A & 0 \\ 0 & -I_m \end{bmatrix} \text{ and } C_0 \coloneqq \begin{bmatrix} 0 & A^T \\ c^T & -b^T \end{bmatrix}.$$

Now, suppose $\hat{z} = (\hat{x}, \hat{y})$ solves (89), (90). Applying Lemma D.5, we conclude that there exists some optimal solution $z^\star$ satisfying (91) such that

$$\begin{aligned} \|z^\star - \hat{z}\|_2 &\leq H(A_0, C_0) \left\| \begin{bmatrix} (A\hat{x} - b)_+ \\ (-\hat{y})_+ \\ A^T\hat{y} - c \\ c^T\hat{x} - b^T\hat{y} \end{bmatrix} \right\|_2 \\ &= H(A_0, C_0) \left\| \begin{bmatrix} \hat{c} - c \\ c^T\hat{x} - \hat{c}^T\hat{x} \end{bmatrix} \right\|_2 \end{aligned}$$

where in the second line, we use the fact that $\hat{z}$ must satisfy the optimality conditions (91) except replacing c with $\hat{c}$. We know by Cauchy-Schwartz that

$$\|z^\star - \hat{z}\|_2 \leq H(A_0, C_0)\|\hat{c} - c\|_2(1 + \|\hat{x}\|_2).$$

Now, the rest of the analysis reduces to bounding $\|\hat{x}\|_2$.

To do this, let $N = \{i : \hat{y}_i = 0\}$ and let $I_N \in \mathbb{R}^{(m-|N|)\times m}$ is the $m \times m$ identity matrix but with the rows corresponding to $N \subset [m]$ deleted, note that the condition $(A\hat{x} - b) \odot \hat{y} = 0$ is equivalent to the condition $I_N(A\hat{x} - b) = 0$ since $A\hat{x} - b \leq 0, \hat{y} \geq 0$. As a result, there exists $N \subset [m]$ such that $\hat{x}$ is optimal if

$$A\hat{x} \leq b \text{ and } I_N A\hat{x} - I_N b = 0.$$

Note that $\hat{x}$ is the minimum norm solution to the primal problem by assumption, so it must be the minimum norm solution to this problem as well (even if the choice of N is not unique). This implies

$$\hat{x} = \arg\min \frac{1}{2}\|x\|_2^2 \text{ s.t. } Ax \leq b, I_N Ax = I_N b.$$

Now, Lemma D.5 implies that there exists some solution $\tilde{x}$ to $Ax \leq b, I_N Ax = I_N b$ such that

$$\|\tilde{x} - 0\|_2 \leq H(A, I_N A) \left\| \begin{bmatrix} (-b)_+ \\ I_N b \end{bmatrix} \right\|_2 \leq 2H(A, I_N A)\|b\|_2 \leq 2\|b\|_2 \max_{N \subset [m]} H(A, I_N A).$$

Of course, since $\hat{x}$ is the minimum norm solution, we have $\|\hat{x}\|_2 \leq \|\tilde{x}\|_2$.

□

D.5 Explicitly bounding the moments of the Hoffman constant

Theorem 3.4 requires the assumption that the scaled Hoffman constant $H(X)$ has at least two moments. We give several justifications for this assumption in Appendices D.3 and D.4, but it is generally hard to formally verify this condition. Indeed, analytical analysis or even mere computation of Hoffman constants is known to be a particularly challenging problem (e.g. Zualinescu, 2003; Ramdas and Peña, 2016). However, in this section, we are able to show that $H(X)$ has two moments as long as a "general position" condition holds on the true conditional PMF.

That said, we emphasize that our analysis in this section is quite conservative; we suspect that $\mathbb{E}[|H(X)|^2] < \infty$ holds in many settings where the general position condition below does not hold.

[12] Interestingly, Hsieh et al. (2022) also use this dual gap condition, although their technical arguments are otherwise unrelated to ours.

Assumption D.2. *Fix any $\mathcal{Y}_0, \mathcal{Y}_1 \subset \mathcal{Y} = \{y_1, \ldots, y_m\}$. Define $\delta(X)$ to be the squared difference between the conditional probabilities that $Y(0) \in \mathcal{Y}_0$ and $Y(1) \in \mathcal{Y}_1$ given X. Formally,*

$$\delta_{\mathcal{Y}_0,\mathcal{Y}_1}(X) = \left(\mathbb{P}(Y(0) \in \mathcal{Y}_0 \mid X) - \mathbb{P}(Y(1) \in \mathcal{Y}_1 \mid X)\right)^2.$$

Define $r_{\mathcal{Y}_0,\mathcal{Y}_1}(X) = \begin{cases} \frac{1}{\delta_{\mathcal{Y}_0,\mathcal{Y}_1}(X)} & \delta_{\mathcal{Y}_0,\mathcal{Y}_1}(X) \neq 0 \\ 0 & \delta_{\mathcal{Y}_0,\mathcal{Y}_1}(X) = 0 \end{cases} < \infty$ to be the generalized reciprocal of $\delta_{\mathcal{Y}_0,\mathcal{Y}_1}(X)$. We assume that there exists $M < \infty$ such that $\mathbb{E}[|r_{\mathcal{Y}_0,\mathcal{Y}_1}(X)|^2] \leq M$ for all $\mathcal{Y}_0, \mathcal{Y}_1 \subset \mathcal{Y}$.

Assumption D.2 requires that for each $\mathcal{Y}_0, \mathcal{Y}_1$, the generalized reciprocal of $[\mathbb{P}(Y(0) \in \mathcal{Y}_0 \mid X) - \mathbb{P}(Y(1) \in \mathcal{Y}_1 \mid X)]^2$ has two moments. This condition is related to the fact that linear programs may become unstable if the angle between two constraint vectors becomes too small (which may happen if $\delta_{\mathcal{Y}_0,\mathcal{Y}_1}(X)$ is small) but are stable if the constraint vectors are perfectly collinear (in which case $\delta_{\mathcal{Y}_0,\mathcal{Y}_1}(X) = r_{\mathcal{Y}_0,\mathcal{Y}_1}(X) = 0$). Since we work with generalized reciprocals, we note that Assumption D.2 automatically holds if $Y(1) \mid X \stackrel{\mathrm{d}}{=} Y(0) \mid X$, in which case $\delta_{\mathcal{Y}_0,\mathcal{Y}_1}(X) = r_{\mathcal{Y}_0,\mathcal{Y}_1}(X) = 0$ a.s.

Proposition D.3. *Suppose $\mathcal{P}$ is the unrestricted class of all distributions, $\mathcal{Y} = \{y_1, \ldots, y_m\}$ is finite, and that $\theta(P) = \mathbb{E}_P[f(Y(1), Y(0))]$. Following the notation in Theorem 3.4, under Assumption D.2, there exists a universal constant C depending only on $|\mathcal{Y}|$ such that $\mathbb{E}[|H(X)|^2] < CM < \infty$.*

Proof sketch. As notation, recall that $H(x)$ is a Lipschitz constant such that

$$\|\nu_x^\star - \hat{\nu}_x\|_2^2 \leq H(x)\|p(x) - \hat{p}(x)\|_2^2.$$

In particular, Lemma 3.3 proves that $H(x) < \infty$ by noting that we can write

$$\hat{\nu}_x \in \operatorname*{arg\,max}_{\nu_x \in \mathbb{R}^{2m}} \nu_x^T \hat{p}(x) \text{ s.t. } A\nu_x \leq c$$

$$\nu_x^\star \in \operatorname*{arg\,max}_{\nu_x \in \mathbb{R}^{2m}} \nu_x^T p^\star(x) \text{ s.t. } A\nu_x \leq c$$

where $c \in \mathbb{R}^{m^2}$ is the concatenation of $\{f(y_0, y_1)\}_{y \in \mathcal{Y}}$ and the optimal transport matrix A can be written as

$$A = \begin{bmatrix} 1_{m\times 1} & 0_{m\times 1} & \cdots & 0_{m\times 1} & I_{m\times m} \\ 0_{m\times 1} & 1_{m\times 1} & \cdots & 0_{m\times 1} & I_{m\times m} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0_{m\times 1} & 0_{m\times 1} & \cdots & 1_{m\times 1} & I_{m\times m} \end{bmatrix} \in \mathbb{R}^{m^2 \times 2m}.$$

Lemma D.6 shows that there exists a universal constant c_1 depending only on $|\mathcal{Y}|$ such that

$$H(x) \leq c_1 H(A_0, C) \text{ for}$$

$$A_0 \coloneqq \begin{bmatrix} A & 0 \\ 0 & -I_{m^2} \end{bmatrix} \text{ and } C_0 \coloneqq \begin{bmatrix} 0 & A^T \end{bmatrix} \text{ and } C = \begin{bmatrix} 0 & A^T \\ p^\star(x)^T & -c^T \end{bmatrix},$$

where $H(A_0, C)$ is the Hoffman constant defined by Hoffman (1952). We note that $H(A_0, C)$ depends on x only through the last row of C, which depends on $p^\star(x)$. To analyze the dependence of $H(A_0, C)$ on x, we have a three-part strategy:

1. Zualinescu (2003) introduce a combinatorial characterization of $H(A_0, C)$. Using this, we prove a general "rank-one update" formula for Hoffman constants. In particular, we bound $H(A_0, C)$ in terms of $H(A_0, C_0)$ and the norm of the residual after projecting $[p^\star(x)^T, -c^T]$ onto the row space of A.
2. We explicitly analyze the eigenstructure of the optimal transport matrix A to bound the residual norm mentioned above.
3. We then combine these results to prove that there exist universal constants c_2, c_3 depending only on $|\mathcal{Y}|$ such that

$$H(x) \leq c_1 H(A_0, C) \leq c_2 + c_3 \max_{\mathcal{Y}_0, \mathcal{Y}_1 \subset \mathcal{Y}} \left(\mathbb{P}(Y(1) \in \mathcal{Y}_1 \mid X = x) - \mathbb{P}(Y(0) \in \mathcal{Y}_0 \mid X = x)\right)^{-1}$$

where above, the power of -1 denotes the generalized reciprocal—in particular, this final result is proved in Lemma D.11. By Assumption D.2, we know that each term in the max above has two moments. Since this is a maximum over finitely many random variables, this implies that $H(X)$ has two moments, as desired. □

D.5.1 Rank one updates for Hoffman constants

Lemma D.7 (Application of Zualinescu (2003) Prop. 5.1)**.** *Suppose $A \in \mathbb{R}^{m\times n}, C \in \mathbb{R}^{\ell\times n}$ and let $H(A,C)$ be the Hoffman constant associated with $\{Ax \le b, Cx = d\}$. Assume C has full row rank (implying $\ell \le n$) and define*

$$\mathcal{K} = \left\{ K \subset [m] : \begin{bmatrix} A_K \\ C \end{bmatrix} \text{ has linearly independent rows} \right\}.$$

Then

$$\begin{aligned} H(A,C)^{-2} &= \min_{K\in\mathcal{K}} \min_{\lambda\in\mathbb{R}^{|K|}_{\ge 0}, v\in\mathbb{R}^{\ell}, \|(\lambda,v)\|_2^2=1} \|A_K^T\lambda + C^T v\|_2^2 \\ &\ge \min_{K\in\mathcal{K}} \min_{\|(\lambda,v)\|_2^2=1} \|A_K^T\lambda + C^T v\|_2^2 \\ &= \min_{K\in\mathcal{K}} \lambda_{\min}\left(\begin{bmatrix} A_K \\ C \end{bmatrix} \begin{bmatrix} A_K \\ C \end{bmatrix}^T \right). \end{aligned}$$

Proof. The first equality follows from Proposition 5.1 of Zualinescu (2003); the rest follows immediately by the definition of an eigenvalue and simple properties of singular values. □

Lemma D.8 (Rank one update for Hoffman constants)**.** *Suppose $A \in \mathbb{R}^{m\times n}, C_0 \in \mathbb{R}^{(\ell-1)\times n}$ where C_0 has full row rank and $\ell \in [n]$. Fix $v \in \mathbb{R}^n$ and let $C = \begin{bmatrix} C_0 \\ v^T \end{bmatrix} \in \mathbb{R}^{\ell\times n}$.*

Define $\mathcal{K}$ to be the subsets of the rows of A such that $\begin{bmatrix} A_K \\ C \end{bmatrix}$ has linearly independent rows. For each $K \in \mathcal{K}$, define $D_K = \begin{bmatrix} A_K \\ C_0 \end{bmatrix} \in \mathbb{R}^{(|K|+\ell-1)\times n}$ and let ϵ_K denote the squared norm of the projection of v onto the orthogonal complement of the row space of D_K, i.e., $\epsilon_K = \|(I_n - D_K^T(D_K D_K^T)^{-1}D_K)v\|_2^2$. Finally, let $\epsilon_0 = \min_{K\in\mathcal{K}} \epsilon_K$.

If $H(A,C)$ is the Hoffman constant associated with the system $\{x : Ax \le b, Cx = d\}$, then there exist universal constants c_0, c_1 depending only on A and C_0 such that

$$H(A,C)^2 \le c_0 + \frac{1 + c_1\|v\|_2^2}{\epsilon_0^2}.$$

where in particular $c_0 = H(A, C_0)$.

Proof. As notation, let $\lambda_{\min\neq 0}(M)$ denote the minimum nonzero eigenvalue of a square matrix M and let $\lambda_k(M)$ denote its kth largest eigenvalue. For each $K \in \mathcal{K}$, let σ_K denote the smallest nonzero singular value of $D_K = \begin{bmatrix} A_K \\ C_0 \end{bmatrix} \in \mathbb{R}^{(|K|+\ell-1)\times n}$.

We assume C_0 is full rank but not C, so there are two cases. In the first case, C is full rank. Then Lemma D.7 gives that

$$\begin{aligned} H(A,C)^{-2} &\ge \min_{K\in\mathcal{K}} \lambda_{\min}\left(\begin{bmatrix} D_K \\ v^T \end{bmatrix} \begin{bmatrix} D_K \\ v^T \end{bmatrix}^T \right) \\ &\ge \min_{K\in\mathcal{K}} \lambda_{\min\neq 0}\left(D_K^T D_K + vv^T \right). \end{aligned}$$

Since $\begin{bmatrix} D_K \\ v^T \end{bmatrix}$ have linearly independent rows, we note that $D_K^T D_K$ has rank $|K|+\ell-1$ and $D_K^T D_K + vv^T$ has rank $|K| + \ell$. This allows us to apply the rank-one eigenvalue perturbation bound from Ipsen and Nadler

(2009), reviewed in Lemma D.9, which implies that

$$\begin{aligned}\lambda_{\min\neq 0}\left(D_K^T D_K + vv^T\right) &= \lambda_{|K|+\ell}\left(D_K^T D_K + vv^T\right)\\ &\geq \frac{1}{2}\left(\sigma_K^2 + \|v\|_2^2 - \sqrt{(\sigma_K^2+\|v\|_2^2)^2 - 4\sigma_K^2\epsilon_K^2}\right)\\ &\geq \frac{1}{2}\left(\sigma_K^2 + \|v\|_2^2 - \sqrt{\left(\sigma_K^2+\|v\|_2^2 - 2\frac{\sigma_K^2\epsilon_K^2}{\sigma_K^2+\|v\|_2^2}\right)^2}\right)\\ &= \frac{\sigma_K^2\epsilon_K^2}{\sigma_K^2+\|v\|_2^2} = \frac{\epsilon_K^2}{1+\|v\|_2^2/\sigma_K^2}\end{aligned}$$

where the last inequality uses the condition that $(\sigma_K^2+\|v\|_2^2)^2 - 4\sigma_K^2\epsilon_K^2 \geq 0$. At this point, note that we can uniformly lower bound σ_K^2 by a strictly positive real number σ_0^2 which does not depend on v. To see this, let $\mathcal{K}' \coloneqq \left\{K \subset [m] : \begin{bmatrix}A_K\\ C_0\end{bmatrix} \text{ has linearly independent rows}\right\}$ and note that $\mathcal{K}'$ does not depend on v since it depends only on C_0, not C. Furthermore, since $\mathcal{K} \subset \mathcal{K}'$ by definition,

$$\min_{K\in\mathcal{K}} \sigma_K^2 \leq \min_{K\in\mathcal{K}'} \sigma_K^2 \coloneqq \sigma_0^2 > 0.$$

σ_0^2 is strictly positive because by definition of $\mathcal{K}'$, each σ_K^2 for $K \in \mathcal{K}'$ is strictly positive, and $\mathcal{K}'$ has finite cardinality. Thus, we can uniformly lower bound σ_K^2 by σ_0^2.

Now, combining the previous results, we observe that

$$H(A,C)^{-2} \geq \frac{\epsilon_0^2}{1+\|v\|_2^2/\sigma_0^2} \Longrightarrow H(A,C)^2 \leq \frac{1+\|v\|_2^2/\sigma_0^2}{\epsilon_0^2}$$

where we remind the reader that $\epsilon_0 \coloneqq \min_{K\in\mathcal{K}} \epsilon_K$.

In the second case, C is not full rank and v can be expressed as a linear combination of the rows of C_0. In this case, for any $b \in \mathbb{R}^m, d \in \mathbb{R}^\ell$, the additional constraint imposed by v either causes $\{Ax \leq b, Cx \leq d\}$ to be empty (which has no effect on the Hoffman constant), or the additional constraint imposed by v is redundant and $\{Ax \leq b, Cx = d\} = \{Ax \leq b, C_0 x \leq d_{1:(\ell-1)}\}$, which also has zero effect on the Hoffman constant. As a result, we conclude that in this case

$$H(A,C)^{-2} = H(A,C_0)^{-2}.$$

Combining the cases yields

$$H(A,C)^2 \leq H(A,C_0)^2 + \frac{1+\|v\|_2^2/\sigma_0^2}{\epsilon_0^2}$$

which concludes the proof. □

Lemma D.9 (Ipsen and Nadler (2009), Corollary 2.7.)**.** *Fix any symmetric matrix $M \in \mathbb{R}^{n\times n}$ and any vector v where M has the eigendecomposition*

$$M = \sum_{i=1}^{k-1} d_i u_i u_i^T \text{ for eigenvalues } d_1 \geq d_2 \geq \ldots d_{k-1} > 0 \text{ and eigenvectors } u_1, \ldots, u_{k-1}.$$

Let $\lambda_k(M+vv^T)$ denote the kth largest eigenvalue of $M+vv^T$, and let ϵ denote the norm of the projection of v onto the orthogonal complement of $u_1, \ldots, u_{k-1}$. Then

$$\lambda_k(M+vv^T) \geq \frac{1}{2}\left[d_{k-1} + \|v\|_2^2 - \sqrt{(d_{k-1}+\|v\|_2^2)^2 - 4d_{k-1}\epsilon^2}\right].$$

D.5.2 Properties of the optimal transport constraint matrix

Lemma D.10 (Properties of the optimal transport constraint matrix)**.** *Fix $m \in \mathbb{N}$ and define*

$$A = \begin{bmatrix} 1_{m\times 1} & 0_{m\times 1} & \cdots & 0_{m\times 1} & I_{m\times m}\\ 0_{m\times 1} & 1_{m\times 1} & \cdots & 0_{m\times 1} & I_{m\times m}\\ \vdots & \vdots & \vdots & \vdots & \vdots\\ 0_{m\times 1} & 0_{m\times 1} & \cdots & 1_{m\times 1} & I_{m\times m}\end{bmatrix} \in \mathbb{R}^{m^2\times 2m}.$$

In other words, the rows of A are simply the row vectors $\{\begin{bmatrix} e_i & e_j \end{bmatrix}\}_{i,j\in[m]}$, where $e_i \in \mathbb{R}^{1\times m}$ denotes the i-th canonical basis vector in $\mathbb{R}^m$. Then the following holds:

1. *Let $\kappa \coloneqq \begin{bmatrix} -1_m & 1_m \end{bmatrix}$. Then the null space of A, denoted $\text{null}(A)$, is simply then span of κ, i.e., $\text{null}(A) = \text{span}(\kappa)$. This implies $\text{rank}(A) = 2m-1$.*
2. *Let $a_1, \dots, a_K$ denote any K linearly independent rows of A. Then there exists $I, J \subset [m]$ such that at least one of I, J is a nonempty strict subset of $[m]$ satisfying the following. For any vector $\mu = [\mu_0, \mu_1] \in \mathbb{R}^{2m}$ such that $\mu_0, \mu_1 \in \mathbb{R}^m$ are probability vectors, let μ_r denote the residual vector after projecting out $a_1, \dots, a_K$ from μ. Then*

$$\|\mu_r\|_2^2 \geq \frac{1}{2m}\left(\sum_{i\in I}\mu_i - \sum_{j\in J}\mu_{j+m}\right)^2$$

   holds for all μ which are linearly independent from $a_1, \dots, a_K$.
3. *In the above result, if $K = 2m-2$, then $\frac{1}{2m}\left(\sum_{i\in I}\mu_i - \sum_{j\in J}\mu_{j+m}\right)^2 > 0$.*

Proof. **First result**. It is easy to see by definition of κ that $\begin{bmatrix} e_i & e_j \end{bmatrix}^T \kappa = 0$ for any $i, j \in [m]$; therefore $\text{span}(\kappa) \subset \text{null}(A)$.

To show that $\text{null}(A) \subset \text{span}(\kappa)$, fix any vector $v \in \mathbb{R}^{2m} \notin \text{span}(\kappa)$. Note that $v \in \text{span}(\kappa)$ if and only if both of the following hold: (a) each entry of v has the same absolute value and (b) $v_{1:m} = -v_{(m+1):2m}$. Since $v \notin \text{span}(\kappa)$ by assumption, either (a) or (b) does not hold. We now deal with these cases in turn.

In case (a), there exist two coordinates $i, j \in [2m]$ such that $|v_i| \neq |v_j|$. Assume WLOG that $i, j \in [m]$ (the proof is analogous even if not); in this case, we can see that

$$\left(\begin{bmatrix} e_i & e_1 \end{bmatrix} - \begin{bmatrix} e_j & e_1 \end{bmatrix}\right)^T v = v_i - v_j \neq 0$$

and thus $v \notin \text{null}(A)$.

In case (b), there exists $i \in [m]$ such that $v_i \neq -v_{i+m}$. Then we observe

$$\left(\begin{bmatrix} e_i & e_1 \end{bmatrix} + \begin{bmatrix} e_1 & e_i \end{bmatrix} - \begin{bmatrix} e_1 & e_1 \end{bmatrix}\right)^T v = v_i + v_{i+m} \neq 0$$

This proves $\text{null}(A) = \text{span}(\kappa)$. By the rank-nullity theorem, this implies $\text{rank}(A) = 2m-1$.

Second result. Suppose that $a_1, \dots, a_K, \mu$ are linearly independent. Note that μ is an element of the row space of A: this is because $\mu^T\kappa = \mu_0^T 1_m - \mu_1^T 1_m = 0$, and thus μ is orthogonal to the null space of A. Since $\text{rank}(A) = 2m-1$, this implies that $K \leq 2m-2$. Also, as notation, the definition of A ensures that we can represent $a_k = \begin{bmatrix} e_{i_k} & e_{j_k} \end{bmatrix}$ for pairs of coordinates $(i_k, j_k) \in [m] \times [m]$ for $k = 1, \dots, K$.

To bound the norm of $\|\mu_r\|_2^2$, we will explicitly find a vector which is orthogonal to $\{\begin{bmatrix} e_{i_k} & e_{j_k} \end{bmatrix}\}_{k\in[K]}$ but does not lie in the span of κ. In particular, suppose that there exists some $I \subset [m], J \subset [m]$ such that (1) I is a nonempty proper subset of $[m]$ and (2) $\{k \in [K] : i_k \in I\} = \{k \in [K] : j_k \in J\}$. In other words, the pairs $\{(i_k, j_k)\}_{k\in K}$ have the relationship that $i_k \in I$ if and only if $j_k \in J$ across all $k \in [K]$. In a moment, we will show that such an I and J exist. For now, we suppose that I, J exist and use them to show the result of the proof.

Given such subsets I, J, define the vector

$$b_{I,J} \coloneqq [e_I \ \ -e_J] \in \mathbb{R}^{2m}$$

where above, as notation, $e_I \coloneqq \sum_{i\in I} e_i$ and $e_J \coloneqq \sum_{j\in J} e_j$. The definition of I, J allows us to easily check that $b_{I,J}$ is orthogonal to $\{\begin{bmatrix} e_{i_k} & e_{j_k} \end{bmatrix}\}_{k\in[K]}$: in particular,

$$\begin{bmatrix} e_{i_k} & e_{j_k} \end{bmatrix}^T b_{I,J} = \begin{cases} 0 & i_k \notin I \text{ and } j_k \notin J \\ 1-1=0 & i_k \in I \text{ and } j_k \in J \end{cases}$$

where the two cases listed above are the *only* two cases by construction of I, J. If μ_r is the projection of μ onto the orthogonal complement of $a_1, \dots, a_K$, this implies that

$$\|\mu_r\|_2^2 \geq \frac{1}{\|b_{I,J}\|_2^2}\left(b_{I,J}^T\mu\right)^2 = \frac{\left(\sum_{i\in I}\mu_i - \sum_{j\in J}\mu_{j+m}\right)^2}{|I|+|J|} \geq \frac{1}{2m}\left(\sum_{i\in I}\mu_i - \sum_{j\in J}\mu_{j+m}\right)^2$$

which is the desired result. As a result, all that is left to prove is the existence of I and J.

To see this, consider the bipartiate graph with vertices $V = \{(v_1, \dots, v_m, w_1, \dots, w_m)\}$ where we say that there is an edge between (v_i, w_j) if and only if $\begin{bmatrix} e_i & e_j \end{bmatrix} \in a_1, \dots, a_K$, and there are no edges among $(v_1, \dots, v_m)$ and $(w_1, \dots, w_m)$. This is a graph with $2m$ vertices and less than $2m-2$ edges, so it cannot be connected, since a connected graph with $2m$ vertices must have at least $2m-1$ edges. Thus, there exist two vertices in V where there is no path between the vertices. Without loss of generality, assume that one of these two vertices is an element of $(v_1, \dots, v_m)$ and call it v_ℓ for some $\ell \in [m]$. Then define

$$I = \{i \in [m] : v_i \text{ is reachable from } v_\ell \text{ or } i = \ell\}$$

$$J = \{j \in [m] : w_j \text{ is reachable from } v_\ell\}$$

Now we must show that (1) $i_k \in I \Leftrightarrow j_k \in J$ and (2) at least one of I, J is a nonempty strict subset of $[m]$.

To show (1), suppose that $i_k \in I$, in other words, there is a path from v_ℓ to v_{i_k}. By definition of the edge set, there is an edge between v_{i_k} and w_{j_k}, and therefore there is a path from v_ℓ to w_{j_k} (with v_{i_k} in the middle). Thus, $j_k \in J$ by definition of J. This proves that $i_k \in I \implies j_k \in J$, and the converse follows immediately from the same logic, proving (1).

To show (2), observe that I is nonempty since it includes ℓ. If I is not a strict subset of $[m]$, that is, $I = [m]$, then J must be a nonempty strict subset of $[m]$. This is because (i) $J \neq [m]$ because otherwise the graph would be fully connected, and (ii) $J \neq \emptyset$ because there are no edges among $v_1, \dots, v_m$, so if $v_1, \dots, v_m$ are connected, they must be connected to at least one of $w_1, \dots, w_m$. This completes the proof of the second result.

Third result. Suppose $K = 2m-2$. Then if $a_1, \dots, a_K, \mu$ are linearly independent, they must span the full row space of A, which has rank $2m-1$ (and note that μ is an element of the row space of A). Now, suppose for sake of contradiction that

$$b_{I,J}^T \mu = \sum_{i \in I} \mu_i - \sum_{j \in J} \mu_{j+m} = 0.$$

Since $\mu^T \kappa = 0$ as well, this implies that μ is orthogonal to $\text{span}(b_{I,J}, \kappa)$. Since $b_{I,J}, \kappa$ are two linearly independent vectors which are both orthogonal to $a_1, \dots, a_K$, and $a_1, \dots, a_K$ have rank $2m-2$, this implies that $\mu \in \text{span}(a_1, \dots, a_K)$, which contradicts the assumption that $a_1, \dots, a_K, \mu$ are linearly independent. □

D.5.3 Putting it all together

Lemma D.11. *Fix $m \in \mathbb{N}$ and define the matrix*

$$A = \begin{bmatrix} 1_{m\times 1} & 0_{m\times 1} & \cdots & 0_{m\times 1} & I_{m\times m} \\ 0_{m\times 1} & 1_{m\times 1} & \cdots & 0_{m\times 1} & I_{m\times m} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0_{m\times 1} & 0_{m\times 1} & \cdots & 1_{m\times 1} & I_{m\times m} \end{bmatrix} \in \mathbb{R}^{m^2 \times 2m}.$$

For any $c \in \mathbb{R}^{m^2}$ and $\mu(x) \in \mathbb{R}^{2m}$ which is the concatenation of two m-length probability vectors, define

$$A_0 \coloneqq \begin{bmatrix} A & 0 \\ 0 & -I_{m^2} \end{bmatrix} \text{ and } C_0 \coloneqq \begin{bmatrix} 0 & [A^T]_{1:2m-1} \end{bmatrix} \text{ and } C = \begin{bmatrix} 0 & [A^T]_{1:2m-1} \\ \mu(x)^T & -c^T \end{bmatrix}.$$

Finally, for any $I, J \subset [m]$, define

$$\delta_{I,J}(x) \coloneqq \left(\sum_{i \in I} \mu_i(x) - \sum_{j \in J} \mu_{j+m}(x) \right)^2.$$

Then there exist universal constants c_0, c_1 depending only on m and c such that

$$H(x)^2 \coloneqq H(A_0, C)^2 \leq c_0 + c_1 \max_{I,J \subset [m]} \frac{\mathbb{I}(\delta_{I,J}(x) = 0)}{\delta_{I,J}(x)} < \infty.$$

where we use the convention that $\frac{0}{0} = 0$, so the right-hand term is always finite. This implies that if X is a random variable such that $\frac{\mathbb{I}(\delta_{I,J}(X)=0)}{\delta_{I,J}(X)}$ has a kth moment for each $I, J \subset [m]$, then

$$\mathbb{E}[|H(X)|^{2k}] < \infty.$$

Proof. Lemma D.10 implies that C_0 is full rank, so we may apply the "Hoffman rank-one update formula" from Lemma D.8. To do this, we need the following notation:

- Define $\mathcal{K}$ to be the subsets of the rows of A_0 such that $\begin{bmatrix} A_K \\ C \end{bmatrix}$ has linearly independent rows.
- For each $K \in \mathcal{K}$, define $D_K = \begin{bmatrix} [A_0]_K \\ C_0 \end{bmatrix}$ and let ϵ_K denote the squared norm of the projection of $[\mu(x)^T, -c^T]$ onto the orthogonal complement of the row space of D_K. We also let $\epsilon_0 = \min_{K \in \mathcal{K}} \epsilon_K$.

Then by Lemma D.8, there exist universal constants c_0, c_1 depending only on A and C_0 (which thus do not depend on x) such that

$$H(A_0, C)^2 \le c_0 + \frac{(1+c_1)\|[\mu(x)^T, -c^T]\|_2^2}{\|\epsilon_0\|_2^2} \le \frac{(1+c_1)(4+\|c\|_2^2)}{\|\epsilon_0\|_2^2}$$

where the above equation uses the fact that $\|\mu(x)\|_1 = 2$ since it is the concatenation of two probability vectors. Since $\|c\|_2^2$ does not change with x, we can reset the values of c_0, c_1 to conclude that

$$H(A_0, C)^2 \le c_0 + \frac{c_1}{\|\epsilon_0\|_2^2}.$$

The only quantity here which depends on $\mu(x)$ is ϵ_0. To analyze its behavior, we must analyze $\{\epsilon_K\}_{K \in \mathcal{K}}$. To do this, we need even more notation. Indeed, for each $K \in \mathcal{K}$, by definition of A_0 and C_0 there exists some $K_1, K_2 \subset [m^2]$ such that

$$D_K \coloneqq \begin{bmatrix} [A_0]_K \\ C_0 \end{bmatrix} = \begin{bmatrix} A_{K_1} & 0 \\ 0 & -[I_{m^2}]_{K_2} \\ 0 & [A^T]_{1:(2m-1)} \end{bmatrix} \coloneqq \begin{bmatrix} A_{K_1} & 0 \\ 0 & B_{K_2} \end{bmatrix}$$

where above, $[I_{m^2}]_{K_2} \in \mathbb{R}^{|K_2| \times m^2}$ selects the rows of I_{m^2} corresponding to the elements of K_2 and B_{K_2} is defined as $B_{K_2} \coloneqq \begin{bmatrix} [I_{m^2}]_{K_2} \\ [A^T]_{1:(2m-1)} \end{bmatrix} \in \mathbb{R}^{(|K_2|+2m-1) \times m^2}$.

Let $r(K) \in \mathbb{R}^{2m+m^2}$ denote the projection of $[\mu(x)^T, -c^T]$ onto the orthogonal complement of the rows of D_K, so $\epsilon_K \coloneqq \|r_K\|_2^2$. The block zeros in D_K ensure that the projections of $\mu(x)$ and c happen *separately*. More precisely, let $r_{\mu(x)}(K_1) \in \mathbb{R}^{2m}$ denote the projection of $\mu(x)$ onto the orthogonal complement of the row span of A_{K_1} and let $r_c(K_2) \in \mathbb{R}^{m^2}$ denote the projection of c onto the orthogonal complement of the row span of B_{K_2}. Then separability yields that

$$r(K) = \begin{bmatrix} r_{\mu(x)}(K_1) \\ r_c(K_2) \end{bmatrix}.$$

Since the rows of D_K and $[\mu(x)^T, -c^T]$ are linearly independent, $\|r(K)\|_2^2 > 0$ and at most one of $r_{\mu(x)}(K_1), r_c(K_2)$ are equal to zero. This implies that *either* A_{K_1} has rows which are linearly independent of $\mu(x)^T$ *or* B_{K_2} has rows which are linearly independent of c^T (but not necessarily both). Thus, if we define

$$\mathcal{K}_1 = \left\{ K_1 \subset [m^2] : \begin{bmatrix} A_{K_1} \\ \mu(x)^T \end{bmatrix} \text{ has linearly independent rows} \right\}$$

$$\mathcal{K}_2 = \left\{ K_2 \subset [m^2] : \begin{bmatrix} B_{K_2} \\ c^T \end{bmatrix} \text{ has linearly independent rows} \right\}$$

we obtain that

$$\|r(K)\|_2^2 \ge \min \left(\min_{K_1 \in \mathcal{K}_1} \|r_{\mu(x)}(K_1)\|_2^2, \underbrace{\min_{K_2 \in \mathcal{K}_2} \|r_c(K_2)\|_2^2}_{\text{does not depend on x}} \right).$$

Note that the outer minimum is a minimum because the definition of $\mathcal{K}_1, \mathcal{K}_2$ ensures that $\min_{K_1 \in \mathcal{K}_1} \|r_{\mu(x)}(K_1)\|_2^2 > 0$ and $\min_{K_2 \in \mathcal{K}_2} \|r_c(K_2)\|_2^2 > 0$—however, as noted above, for any $K \in \mathcal{K}$, we can only ensure that *either* $K_1 \in \mathcal{K}_1$ *or* $K_2 \in \mathcal{K}_2$, not both.

Now, we note that the quantity $\min_{K_2 \in \mathcal{K}_2} \|r_c(K_2)\|_2^2$ does not depend on x and is strictly positive because it is a minimum of finitely many strictly positive real numbers. Therefore, it suffices to bound

$\min_{K_1 \in \mathcal{K}_1} \|r_{\mu(x)}(K_1)\|_2^2$. However, Lemma D.10 does precisely this task. In particular, Lemma D.10 directly implies that if

$$\delta_{I,J}(x) \coloneqq \left(\sum_{i \in I} \mu_i(x) - \sum_{j \in J} \mu_{j+m}(x) \right)^2$$

then

$$\min_{K_1 \in \mathcal{K}_1} \|r_{\mu(x)}(K_1)\|_2^2 \geq \frac{1}{2m} \min_{I,J \subset [m]: \delta_{I,J}(x) > 0} \delta_{I,J}(x).$$

Combining these results, we obtain that there exist universal constants c_2, c_3 depending only on m such that

$$H(A_0, C)^2 \leq c_2 + c_3 \max_{I,J \subset [m]} \frac{\mathbb{I}(\delta_{I,J}(x) = 0)}{\delta_{I,J}(x)}$$

where above we use the convention that $\frac{0}{0} = 0$. This completes the proof. □

D.6 Main proofs from Section 3.3

Proposition 3.1. *Assume the conditions of either Theorem 3.4 or Proposition D.2. Then*

$$\sqrt{n}(\hat{\theta}_L^{\text{crossfit}} - \hat{\theta}_L^{\star\star}) \xrightarrow{\text{p}} 0 \text{ and } \sqrt{n}(\hat{\theta}_{\text{LCB}}^{\text{crossfit}} - \hat{\theta}_{\text{LCB}}^{\star\star}) \xrightarrow{\text{p}} 0.$$

Proof. The proof is identical to the proof of Corollary 3.4. □

Proposition 3.2. *Assume unconfoundedness, strict overlap, and that $\hat{\nu}$ satisfies the moment condition in Assumption 3.2. Assume that $\hat{\nu}^{\text{swap}}$ is computed using the same procedure as $\hat{\nu}$ (but applied to $\mathcal{D}_2$ instead of $\mathcal{D}_1$), so that all assumptions applying to $\hat{\nu}$ apply to $\hat{\nu}^{\text{swap}}$ by symmetry. Finally, assume that one of the following holds:*

1. *Condition 1: There exist arbitrary deterministic functions $\nu_k^\dagger : \mathcal{Y} \times \mathcal{X} \to \mathbb{R}$ satisfying the moment condition in Assumption 3.2 such that $\mathbb{E}\left[\left(\hat{\nu}_k(Y(k), X) - \nu_k^\dagger(Y(k), X)\right)^2\right] \to 0$ holds at any rate for $k \in \{0, 1\}$.*
2. *Condition 2: The outcome model is sufficiently misspecified such that the first-stage bias is larger than $n^{-1/2}$, i.e., $\theta_L - \tilde{\theta}_L = \omega_p(n^{-1/2})$.*

Then

$$\liminf_{n \to \infty} \mathbb{P}(\hat{\theta}_{\text{LCB}}^{\text{crossfit}} \leq \theta_L) \geq 1 - \alpha.$$

Proof. We handle the two conditions separately.

Proof under Condition 1: We first introduce some notation. Define the summand

$$S_i = \begin{cases} \frac{\hat{\nu}_1^{\text{swap}}(Y_i, X_i) W_i}{\pi(X_i)} + \frac{\hat{\nu}_0^{\text{swap}}(Y_i, X_i)(1 - W_i)}{1 - \pi(X_i)} & i \in \mathcal{D}_1 \\ \frac{\hat{\nu}_1(Y_i, X_i) W_i}{\pi(X_i)} + \frac{\hat{\nu}_0(Y_i, X_i)(1 - W_i)}{1 - \pi(X_i)} & i \in \mathcal{D}_2 \end{cases}$$

so that by definition,

$$\hat{\theta}_L = \frac{1}{|\mathcal{D}_2|} \sum_{i \in \mathcal{D}_2} S_i \text{ and } \hat{\theta}_L^{\text{swap}} = \frac{1}{|\mathcal{D}_1|} \sum_{i \in \mathcal{D}_1} S_i \text{ and } \hat{\theta}_L^{\text{crossfit}} = \bar{S} = \frac{\hat{\theta}_L + \hat{\theta}_L^{\text{swap}}}{2}.$$

and $\hat{\theta}_{\text{LCB}}^{\text{crossfit}} = \hat{\theta}_L^{\text{crossfit}} - \Phi^{-1}(1 - \alpha) \frac{\hat{\sigma}_s}{\sqrt{n}}$, where throughout this proof, $\hat{\sigma}_s$ is the sample standard deviation of $\{S_i\}_{i=1}^n$.

Now, we will compare $\hat{\theta}_{\text{LCB}}$ to an oracle estimator. Define $S_i^\dagger$ to be the analogue of S_i, but replacing $\hat{\nu}$ and $\hat{\nu}^{\text{swap}}$ with $\nu^\dagger$:

$$S_i^\dagger \coloneqq \frac{\nu_1^\dagger(Y_i, X_i) W_i}{\pi(X_i)} + \frac{\nu_0^\dagger(Y_i, X_i)(1 - W_i)}{1 - \pi(X_i)}$$

and the oracle estimator and lower confidence bound are defined as

$$\hat{\theta}_L^\dagger \coloneqq \frac{1}{n}\sum_{i=1}^n S_i^\dagger \text{ and } \hat{\theta}_{\text{LCB}}^\dagger \coloneqq \hat{\theta}_L^\dagger - \Phi^{-1}(1-\alpha)\frac{\hat{\sigma}_s^\dagger}{\sqrt{n}}$$

where $\hat{\sigma}_s^\dagger$ is the sample standard deviation of $\{S_i^\dagger\}_{i=1}^n$.

$\hat{\theta}_{\text{LCB}}^\dagger$ is clearly a valid $1-\alpha$ lower confidence bound on $\mathbb{E}[S_i^\dagger]$ by the univariate central limit theorem; thus, by weak duality, $\hat{\theta}_{\text{LCB}}^\dagger$ is a valid $1-\alpha$ lower confidence bound on θ_L (see Theorem 3.1). Thus, a standard proof technique in the literature is to show that $\hat{\theta}_{\text{LCB}}^{\text{crossfit}}$ is asymptotically equivalent to $\hat{\theta}_{\text{LCB}}^\dagger$. However, this is *not* true in this setting: in general, $\sqrt{n}(\hat{\theta}_{\text{LCB}}^\dagger - \hat{\theta}_{\text{LCB}}^{\text{crossfit}}) \not\xrightarrow{p} 0$. Nonetheless, a careful application of weak duality will allow us to show the desired result. In particular, $\hat{\theta}_{\text{LCB}}^{\text{crossfit}}$ may have more fluctuations than $\hat{\theta}_{\text{LCB}}^\dagger$, but weak duality will guarantee that $\hat{\theta}_{\text{LCB}}^{\text{crossfit}}$ will not fluctuate above θ_L with probability more than α asymptotically.

In particular, Lemma D.12 proves the standard result that

$$\begin{aligned}\hat{\theta}_{\text{LCB}}^{\text{crossfit}} - \hat{\theta}_{\text{LCB}}^\dagger &= \frac{\mathbb{E}[\hat{\theta}_L \mid \mathcal{D}_1] + \mathbb{E}[\hat{\theta}_L^{\text{swap}} \mid \mathcal{D}_2]}{2} - \mathbb{E}[S_i^\dagger] + o_p(n^{-1/2}) \\ &= \frac{g(\hat{\nu}) + g(\hat{\nu}^{\text{swap}})}{2} - g(\nu^\dagger) + o_p(n^{-1/2}).\end{aligned} \tag{92}$$

where g is the Kantorovich dual function defined in Section 2, and the latter equality follows from the fact that conditional on $\mathcal{D}_1$, $\hat{\theta}_L$ is simply an IPW estimator for $g(\hat{\nu})$ (and analogously for $\hat{\theta}_L^{\text{swap}}$ and $\hat{\nu}^{\text{swap}}$). Now, $\frac{g(\hat{\nu})+g(\hat{\nu}^{\text{swap}})}{2} - g(\nu^\dagger)$ in general may have fluctuations on a scale larger than $n^{-1/2}$. However, the magic comes from weak duality, which yields that $g(\hat{\nu}), g(\hat{\nu}^{\text{swap}}) \le \theta_L$. As a result, we have that

$$\hat{\theta}_{\text{LCB}}^{\text{crossfit}} - \hat{\theta}_{\text{LCB}}^\dagger \le \theta_L - g(\nu^\dagger) + o_p(n^{-1/2}). \tag{93}$$

Plugging this in, we can show the key validity result:

$$\begin{aligned}\mathbb{P}(\hat{\theta}_{\text{LCB}}^{\text{crossfit}} \ge \theta_L) &= \mathbb{P}(\hat{\theta}_{\text{LCB}}^{\text{crossfit}} - \hat{\theta}_{\text{LCB}}^\dagger \ge \theta_L - \hat{\theta}_{\text{LCB}}^\dagger) \\ &\le \mathbb{P}(\theta_L - g(\nu^\dagger) + o_p(n^{-1/2}) \ge \theta_L - g(\nu^\dagger) + g(\nu^\dagger) - \hat{\theta}_{\text{LCB}}^\dagger) && \text{by Eq. (93)} \\ &= \mathbb{P}(o_p(n^{-1/2}) \ge g(\nu^\dagger) - \hat{\theta}_{\text{LCB}}^\dagger) && \text{by cancellation} \\ &= \mathbb{P}(o_p(n^{-1/2}) \ge \mathbb{E}[S_i^\dagger] - \hat{\theta}_{\text{LCB}}^\dagger) && \text{since } g(\nu^\dagger) = \mathbb{E}[S_i^\dagger].\end{aligned}$$

At this point, the result now follows by applying the standard univariate central limit theorem to $\hat{\theta}_{\text{LCB}}^\dagger$. Formally, we note that we assume that $\nu^\dagger$ satisfies the moment condition in Assumption 3.2, and therefore the argument in Theorem 3.1 proves that we can apply the Lyapunov CLT to $\{S_i^\dagger\}_{i\in[n]}$. Furthermore, Assumption 3.2 directly implies that $\text{Var}(S_i^\dagger)$ is uniformly bounded away form zero; therefore, the CLT implies that there exists some $c > 0$ such that $\mathbb{P}(\mathbb{E}[S_i^\dagger] - \hat{\theta}_{\text{LCB}}^\dagger \ge c/\sqrt{n}) = 1-\alpha-\gamma$. Since $o_p(n^{-1/2}) \ll c/\sqrt{n}$ as $n \to \infty$ by definition, we have that for every $\gamma > 0$,

$$\limsup_n \mathbb{P}(o_p(n^{-1/2}) \ge \mathbb{E}[S_i^\dagger] - \hat{\theta}_{\text{LCB}}^\dagger) \le \alpha + \gamma.$$

Since this holds for all $\gamma > 0$, we conclude that

$$\limsup_n \mathbb{P}(\hat{\theta}_{\text{LCB}}^{\text{crossfit}} \ge \theta_L) \le \limsup_n \mathbb{P}(o_p(n^{-1/2}) \ge \mathbb{E}[S_i^\dagger] - \hat{\theta}_{\text{LCB}}^\dagger) \le \alpha$$

which proves the result.

Proof under Condition 2: Under this condition, a similar proof as Lemma D.2 shows that

$$\hat{\theta}_L = g(\hat{\nu}) + O_p(n^{-1/2}), \hat{\theta}_L^{\text{swap}} = g(\hat{\nu}^{\text{swap}}) + O_p(n^{-1/2}) \text{ and } \hat{\sigma}_s = O_p(1).$$

As a result, we have that

$$\hat{\theta}_{\text{LCB}}^{\text{crossfit}} = \frac{g(\hat{\nu}) + g(\hat{\nu}^{\text{swap}})}{2} + O_p(n^{-1/2}).$$

Condition 2 tells us that $\theta_L - \tilde{\theta}_L \coloneqq \theta_L - g(\hat{\nu}) = \omega_p(n^{-1/2})$, and the same result holds with $g(\hat{\nu}^{\text{swap}})$ replacing $g(\hat{\nu})$ by symmetry. Therefore

$$\hat{\theta}_{\text{LCB}}^{\text{crossfit}} = \theta_L + \underbrace{\frac{g(\hat{\nu}) + g(\hat{\nu}^{\text{swap}})}{2} - \theta_L}_{\omega_p(n^{-1/2})} + O_p(n^{-1/2}).$$

Since the $\omega_p(n^{-1/2})$ term is deterministically nonnegative by weak duality and dominates the $O_p(n^{-1/2})$ term, we conclude

$$\liminf_n \mathbb{P}(\hat{\theta}_{\text{LCB}}^{\text{crossfit}} \leq \theta_L) = 1 \geq 1 - \alpha$$

which proves the desired result. □

Lemma D.12. *Assume the conditions and notation of Proposition 3.2. Then*

$$\hat{\theta}_{\text{LCB}}^{\text{crossfit}} - \hat{\theta}_{\text{LCB}}^{\dagger} = \frac{g(\hat{\nu}) + g(\hat{\nu}^{\text{swap}})}{2} - g(\nu^{\dagger}) + o_p(n^{-1/2}).$$

Proof. First, we observe that

$$\begin{aligned}\hat{\theta}_{\text{LCB}}^{\text{crossfit}} - \hat{\theta}_{\text{LCB}}^{\dagger} &= \hat{\theta}_{L}^{\text{crossfit}} - \hat{\theta}_{L}^{\dagger} + \frac{\Phi^{-1}(1-\alpha)}{\sqrt{n}}\left[\hat{\sigma}_s - \hat{\sigma}_s^{\dagger}\right] \\ &= \hat{\theta}_{L}^{\text{crossfit}} - \hat{\theta}_{L}^{\dagger} + o_p(n^{-1/2})\end{aligned}$$

where the last line follows because $\hat{\sigma}_s - \hat{\sigma}_s^{\dagger} = o_p(1)$ by the same argument as in Lemma D.1.

Next, we observe

$$\hat{\theta}_{L}^{\text{crossfit}} - \hat{\theta}_{L}^{\dagger} = \frac{1}{2|\mathcal{D}_1|}\sum_{i\in\mathcal{D}_1} S_i - S_i^{\dagger} + \frac{1}{2|\mathcal{D}_2|}\sum_{i\in\mathcal{D}_2} S_i - S_i^{\dagger}.$$

For simplicity, we focus on the first sum above. Note that $\{S_i - S_i^{\dagger}\}_{i\in\mathcal{D}_1}$ are i.i.d. conditional on $\mathcal{D}_1$. We will apply Chebyshev's inequality to this sum; to do this, we analyze its mean and variance.

1. *Mean*: Since $\pi(X_i)$ are known propensity scores, we have the exact result that

$$\mathbb{E}[S_i \mid \mathcal{D}_2] - \mathbb{E}[S_i^{\dagger} \mid \mathcal{D}_2] = g(\hat{\nu}^{\text{swap}}) - g(\nu^{\dagger}).$$

2. *Variance*: Observe that

$$\begin{aligned}\text{Var}(S_i - S_i^{\dagger} \mid \mathcal{D}_2) &\leq \mathbb{E}[(S_i - S_i^{\dagger})^2 \mid \mathcal{D}_2] \\ &= \mathbb{E}\left[\left(\frac{W_i(\hat{\nu}_1^{\text{swap}}(X_i,Y_i) - \nu_1^{\dagger}(X_i,Y_i))}{\pi(X_i)}\right)^2 \mid \mathcal{D}_2\right] \\ &+ \mathbb{E}\left[\left(\frac{(1-W_i)(\hat{\nu}_0^{\text{swap}}(X_i,Y_i) - \nu_0^{\dagger}(X_i,Y_i))}{1-\pi(X_i)}\right)^2 \mid \mathcal{D}_2\right] \\ &\leq \Gamma^{-2}\sum_{k\in\{0,1\}} \mathbb{E}[(\hat{\nu}_k^{\text{swap}}(X_i,Y_i) - \nu_k^{\dagger}(X_i,Y_i))^2 \mid \mathcal{D}_2].\end{aligned}$$

   However, we assume that $\mathbb{E}[(\hat{\nu}_k^{\text{swap}}(X_i,Y_i) - \nu_k^{\dagger}(X_i,Y_i))^2] \to 0$ for $k \in \{0,1\}$. Thus, $\mathbb{E}[\text{Var}(S_i - S_i^{\dagger} \mid \mathcal{D}_2)] \to 0$ and thus $\text{Var}(S_i - S_i^{\dagger} \mid \mathcal{D}_2) = o_p(1)$.

From this analysis, we conclude by Chebyshev's inequality that

$$\frac{1}{2|\mathcal{D}_1|}\sum_{i\in\mathcal{D}_1} S_i - S_i^{\dagger} = \frac{g(\hat{\nu}^{\text{swap}}) - g(\nu^{\dagger})}{2} + o_p(n^{-1/2}).$$

The same analysis applied to the other sum yields that

$$\frac{1}{2|\mathcal{D}_2|}\sum_{i\in\mathcal{D}_2} S_i - S_i^{\dagger} = \frac{g(\hat{\nu}) - g(\nu^{\dagger})}{2} + o_p(n^{-1/2}).$$

Combining all of the above results yields the desired result

$$\begin{aligned}\hat{\theta}_{\text{LCB}}^{\text{crossfit}} - \hat{\theta}_{\text{LCB}}^{\dagger} &= \hat{\theta}_{L}^{\text{crossfit}} - \hat{\theta}_{L}^{\dagger} + o_p(n^{-1/2}) \\ &= \frac{g(\hat{\nu}^{\text{swap}}) + g(\hat{\nu})}{2} - g(\nu^{\dagger}) + o_p(n^{-1/2}).\end{aligned}$$

□

E Proof of Counterexample 1

Theorem E.1 (Counterexample 1)**.** *Suppose the vectors $(X_i, W_i, Y_i(0), Y_i(1))$ are sampled i.i.d. from some population distribution $P^\star \in \mathcal{P}$, where $\mathcal{P}$ is the set of distributions on $\mathcal{X} \times \{0,1\} \times \mathcal{Y}^2$ satisfying unconfoundedness, i.e., $\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp W_i \mid X_i$, and strict overlap, i.e., $\inf_{x\in\mathcal{X}} \pi_P(x) > 0$ and $\sup_{x\in\mathcal{X}} \pi_P(x) < 1$ for all $P \in \mathcal{P}$, where $\pi_P(x) \coloneqq \mathbb{E}_P[W \mid X = x]$. We denote the observed response as $Y_i = W_i Y_i(1) + (1-W_i)Y_i(0)$.*

Given i.i.d. observations $\{(X_i, W_i, Y_i)\}_{i=1}^n$, we seek to form a lower bound on the average treatment effect $\theta(P) \coloneqq \mathbb{E}_P[Y(1) - Y(0)]$ which is valid even under arbitrary misspecification of $\pi(X_i)$ and the outcome model. Although θ is identifiable, it can still be written as the solution to the optimization problem

$$\theta(P^\star) = \min_{P\in\mathcal{P}} \mathbb{E}[Y_i(1) - Y_i(0)] \text{ s.t. } P_{X,W,Y} = P^\star_{X,W,Y} \tag{94}$$

where $P_{X,W,Y}$ is the law of (X_i, W_i, Y_i) under P and $P^\star_{X,W,Y}$ is the true (identifiable) law of (X, W, Y). Note that the optimization variable is P, which is a joint law over $(X, W, Y(0), Y(1))$, and $P_{X,W,Y}$ is a functional of P. For any function $h : \mathcal{X} \times \{0,1\} \times \mathcal{Y} \to \mathbb{R}$, the Lagrange dual to this problem is

$$g(h) \coloneqq \mathbb{E}_{P^\star}\left[h(X_i, W_i, Y_i)\right] + \kappa(h),$$

where $\kappa(h) \coloneqq \inf_{P\in\mathcal{P}} \mathbb{E}_P[Y_i(1) - Y_i(0) - h(X_i, W_i, Y_i)]$ is a known constant depending on h. For any h, we have that

$$h(X_i, W_i, Y_i) + \kappa(h) \le \begin{cases} Y_i - \max(\mathcal{Y}) & W_i = 1 \\ \min(\mathcal{Y}) - Y_i & W_i = 0. \end{cases}$$

Proof. The proof is in three steps. First, we review the derivation of the Lagrange dual. Second, we show the result in the setting where there are no covariates; then, we generalize to the case with covariates.

First, we review the form of the Lagrange dual, following Boyd and Vandenberghe (2004). Note that the objective function is the map $o : \mathcal{P} \to \mathbb{R}$ where $P \mapsto \mathbb{E}_P[Y(1) - Y(0)]$ with domain $\mathcal{P}$. The optimization variable is $P \in \mathcal{P}$, a distribution over $(X, W, Y(0), Y(1))$, which induces a distribution $P_{X,W,Y}$ over (X, W, Y). For every $P \in \mathcal{P}$ satisfying strong ignorability and strict overlap, $\mathbb{E}_P[Y(1) - Y(0)]$ is a functional of $P_{X,W,Y}$; thus, $\theta(P^\star)$ is identifiable, and we have the equation

$$\theta(P^\star) = \min_{P\in\mathcal{P}} \mathbb{E}[Y_i(1) - Y_i(0)] \text{ s.t. } P_{X,W,Y} = P^\star_{X,W,Y}.$$

Since our constraint is $P_{X,W,Y} = P^\star_{X,W,Y}$, the Lagrangian is simply the objective function plus an additional linear functional of the difference between $P_{X,W,Y}$ and $P^\star_{X,W,Y}$. In other words, for any $h : \mathcal{X} \times \{0,1\} \times \mathcal{Y} \to \mathbb{R}$, the Lagrangian is defined as

$$L(P, h) = \mathbb{E}_P[Y(1) - Y(0)] + \mathbb{E}_{P^\star}[h(X, W, Y)] - \mathbb{E}_P[h(X, W, Y)] = \mathbb{E}_{P^\star}[h(X, W, Y)] + \mathbb{E}_P[Y(1) - Y(0) - h(X, W, Y)]. \tag{95}$$

The Lagrange dual function is simply the infimum of $L(P, h)$ over $P \in \mathcal{P}$:

$$g(h) = \mathbb{E}_{P^\star}[h(X, W, Y)] + \inf_{P\in\mathcal{P}} \mathbb{E}_P[Y(1) - Y(0) - h(X, W, Y)] = \mathbb{E}_{P^\star}[h(X, W, Y)] + \kappa(h) \tag{96}$$

for $\kappa(h)$ as defined previously, which shows the result.

Now, we show the main result in the setting where $\mathcal{X}$ contains one element and thus there are no covariates. In this case, fix any function $h : \{0,1\} \times \mathcal{Y} \to \mathbb{R}$. For $w, y \in \{0,1\} \times \mathcal{Y}$, we can write

$$h(w, y) = w h_1(y) + (1 - w) h_0(y)$$

for $h_1, h_0 : \mathcal{Y} \to \mathbb{R}$. Note that under any $P \in \mathcal{P}$, we have by weak duality and unconfoundedness that

$$\begin{aligned} \theta(P) &= \mathbb{E}_P[Y(1) - Y(0)] \\ &\ge \mathbb{E}_P[W h_1(Y(1)) + (1 - W) h_0(Y(0)) + \kappa(h)] \\ &= P(W = 1)\mathbb{E}_P[h_1(Y(1))] + P(W = 0)\mathbb{E}_P[h_0(Y(0))] + \kappa(h). \end{aligned}$$

Taking limits as $P(W = 1) \to 1$ and $P(W = 0) \to 0$ (note this does not violate strict overlap for each P as $P(W = 1) \in (0, 1)$), we obtain that

$$\kappa(h) \le \min_{P\in\mathcal{P}} \mathbb{E}_P[Y(1) - Y(0)] - \mathbb{E}_P[h_1(Y(1))].$$

Letting P by any distribution such that $Y(0) = \max(\mathcal{Y})$ with probability one, we obtain

$$\kappa(h) \leq \min_{P \in \mathcal{P}: Y(0) = \max(\mathcal{Y}) \text{ a.s.}} \mathbb{E}_P[Y(1) - h_1(Y(1))] - \max(\mathcal{Y}).$$

We can choose P to be a point mass such that $Y(0) = \max(\mathcal{Y})$ and $Y(1) = \min_{y \in \mathcal{Y}} y - h_1(y)$ which yields $\kappa(h) \leq -\max(\mathcal{Y}) + \min_{y \in \mathcal{Y}} y - h_1(y)$. This directly implies that for any $y \in \mathcal{Y}$,

$$\kappa(h) \leq -\max(\mathcal{Y}) + y - h_1(y) \implies h_1(y) + \kappa(h) \leq y - \max(\mathcal{Y}).$$

Repeating this analysis yields

$$h_0(y) + \kappa(h) \leq \min(\mathcal{Y}) - y$$

which by definition of h_1, h_0 completes the proof in the case where $\mathcal{X}$ has one element. In particular, we proved that if $\mathbb{E}_P[h(W, Y)] \leq \theta(P)$ for all $P \in \mathcal{P}$, then

$$h(W, Y) + \kappa(h) \leq \begin{cases} Y - \max(\mathcal{Y}) & W = 1 \\ \min(\mathcal{Y}) - Y & W = 0. \end{cases}$$

Now consider the general case where $\mathcal{X}$ may have multiple or infinitely many elements. Note that we must have that $\mathbb{E}_P[h(X_i, W_i, Y_i)] + \kappa(h)] \leq \theta(P)$ holds for all $P \in \mathcal{P}$. As a result, this must also hold conditional on $X = x$ for all $x \in \mathcal{X}$ and all $P \in \mathcal{P}$; otherwise, we could consider some P which guarantees that $X = x$ with probability one for some worst case choice of x. Since this must hold conditional on X, it reduces to the case with no covariates. This completes the proof. □